UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO DE BATELADAS PRODUTIVAS

Porto Alegre
2013

Alessandro Kahmann

Seleção de variáveis para classificação de bateladas produtivas

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Qualidade.

Orientador: Michel Jose Anzanello, *Ph.D.*

Porto Alegre

2013

Alessandro Kahmann

Seleção de variáveis para classificação de bateladas produtivas

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

Prof. Michel Jose Anzanello, *Ph.D.*
Orientador PPGEP/UFRGS

_____

Prof. José Luis Duarte Ribeiro, Dr.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professora Liane Werner, Dr. (PPGEP/UFRGS)
Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)
Professora Lisiane Priscila Roldão Selau, Dr. (DEST/UFRGS)

AGRADECIMENTOS

Aproveito este espaço para agradecer a todos que estiveram comigo ao longo dos anos e contribuíram para a chegada deste momento.

A minha esposa Lisiane pelo apoio incondicional, não apenas nestes últimos dois anos, mas desde o momento em que nos conhecemos.

Aos meus pais por me proporcionarem condições para que eu alcançasse meus objetivos.

As minhas irmãs pela companhia ao longo de toda a vida.

Aos meus sogros e cunhados que me receberam de braços abertos em sua família.

Ao meu orientador, Prof. Michel José Anzanello, *Ph.D*., pela oportunidade, apoio e confiança.

Aos meus amigos de infância, Cássio e Gustavo.

Aos meus amigos Fábio, Luiz Ernesto, Nash, Lucas, Bruno, Luís, Petisco, Guilherme e Hélio pela convivência desde meus primeiros anos de faculdade.

Aos professores e colegas do Programa de Pós-Graduação em Engenharia de Produção, em especial ao João, Juliano, Rodolfo, Ricardo e Vitor que tanto meu auxiliaram neste momento de minha vida.

KAHMANN, Alessandro ***Seleção de variáveis para classificação de bateladas produtivas,*** 2013. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Brasil.

RESUMO

Bancos de dados oriundos de processos industriais são caracterizados por elevado número de variáveis correlacionadas, dados ruidosos e maior número de variáveis do que observações, tornando a seleção de variáveis um importante problema a ser analisado no monitoramento de tais processos. A presente dissertação propõe sistemáticas para seleção de variáveis com vistas à classificação de bateladas produtivas. Para tanto, sugerem-se novos métodos que utilizam Índices de Importância de Variáveis para eliminação sistemática de variáveis combinadas a ferramentas de classificação; objetiva-se selecionar as variáveis de processo com maior habilidade discriminante para categorizar as bateladas em classes. Os métodos possuem uma sistematização básica que consiste em: i) separar os dados históricos em porções de treino e teste; ii) na porção de treino, gerar um Índice de Importância de Variáveis (IIV) que ordenará as variáveis de acordo com sua capacidade discriminante; iii) a cada iteração, classificam-se as amostras da porção de treino e removem-se sistematicamente as variáveis; iv) avaliam-se então os subconjuntos através da distância Euclidiana dos resultados dos subconjuntos a um ponto hipotético ótimo, definindo assim o subconjunto de variáveis a serem selecionadas. Para o cumprimento das etapas acima, são testadas diferentes ferramentas de classificação e IIV. A aplicação dos métodos em bancos reais e simulados verifica a robustez das proposições em dados com distintos níveis de correlação e ruído.

Palavras-chave: Seleção de Variáveis; Índice de Importância de Variáveis; Classificação

KAHMANN, Alessandro ***Seleção de variáveis para classificação de bateladas produtivas,*** 2013. Dissertação (Mestrado em Engenharia) - Universidade Federal do Rio Grande do Sul, Brasil.

ABSTRACT

Databases derived from industrial processes are characterized by a large number of correlated, noisy variables and more variables than observations, making of variable selection an important issue regarding process monitoring. This thesis proposes methods for variable selection aimed at classifying production batches. For that matter, we propose new methods that use Variable Importance Indices for variable elimination combined with classification tools; the objective is to select the process variables with the highest discriminating ability to categorize batch classes. The methods rely on a basic framework: i) split historical data into training and testing sets; ii) in the training set, generate a Variable Importance Index (VII) that will rank the variables according to their discriminating ability; iii) at each iteration, classify samples from the training set and remove the variable with the smallest VII; iv) candidate subsets are then evaluated through the Euclidean distance to a hypothetical optimum, selecting the recommended subset of variables. The aforementioned steps are tested using different classification tools and VII's. The application of the proposed methods to real and simulated data corroborates the robustness of the propositions on data with different levels of correlation and noise.

Keywords: Variable Selection; Variable Importance Index; Classification

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. Introdução

1.1 Considerações Iniciais

Processos industriais podem apresentar centenas ou até milhares de variáveis ruidosas ou fortemente correlacionadas. Tendo em vista a importância do monitoramento e controle da qualidade do produto, é de suma importância elaborar modelos capazes de identificar as variáveis que melhor descrevem a qualidade do produto. Em processos que operam em bateladas tal necessidade se acentua, visto que o número de amostras é tipicamente menor que o de variáveis, aumentado a dificuldade de análise uma vez que diversas ferramentas falham sob estas condições. Em tais cenários, torna-se necessária a utilização de técnicas de redução de dimensionalidade que permitam executar a análise dos dados sem a perda de informações relevantes (MARTIN *et al.,* 1999; GAUCHI; CHAGNON, 2001; ANZANELLO *et al.*, 2012).

A mineração de dados é o processo computacional para identificação de padrões dentro de grandes bancos de dados, tendo como principal objetivo extrair informações relevantes destes bancos. Dentre as técnicas de mineração de dados, destaca-se a seleção de variáveis, a qual objetiva selecionar as variáveis mais importantes para a realização da análise, removendo variáveis irrelevantes ou que prejudiquem a interpretação dos dados. A seleção de variáveis pode ser justificada pelos seguintes aspectos: i) evitar o *overfitting* de modelos; ii) produzir modelos com menor necessidade de processamento e melhor custo-efetividade; iii) ter um conhecimento aprofundado do processo, uma vez que a identificação de variáveis com base no conhecimento empírico de especialistas é frequentemente sujeita a equívocos (BLUM; LANGLEY, 1997; GUYON; ELISSEEFF, 2003; HASTIE *et al*., 2005; KETTANEH *et al.*, 2005; SAEYS, 2007; ANZANELLO, 2009)

Em cenários industriais, bancos de dados reduzidos são almejados para viabilizar o monitoramento de parâmetros do processo produtivo, permitindo identificar previamente mudanças de comportamento do processo. Dentro deste contexto, sistemáticas de seleção de variáveis dividem-se em dois objetivos principais: i) predição, em que o objetivo é definir o subconjunto das variáveis independentes que viabilizam a melhor predição do valor de uma ou mais variáveis dependentes, como em Gauchi e Chagnon (2001) e Pereira *et al*. (2011); e ii) classificação, em que o objetivo é definir o subconjunto das variáveis independentes com a melhor habilidade de categorização de amostras, como em Brodnjak-Voncina *et al*. (2005) e Tian *et al*. (2013). As proposições desta dissertação estão alinhadas com o segundo objetivo.

Esta dissertação é composta por três artigos que abordam sistemáticas de seleção de variáveis para classificação de bateladas produtivas. No primeiro artigo é proposto um método de seleção de variáveis a partir da eliminação *backward* de variáveis; as variáveis são ordenadas através de um Índice de Importância de Variáveis (IIV) e as bateladas classificadas pela Distância de Mahalanobis. Seus resultados são comparados aos da regressão *stepwise*, técnica popularmente utilizada para seleção de variáveis. O segundo artigo traz uma variação do método proposto no primeiro artigo através da incorporação da ferramenta de classificação Máquina de Suporte Vetorial. Este método é comparado à eliminação *backward* realizada pela sistemática "Omita Uma Variável por Vez". No terceiro artigo comparam-se combinações de dois novos IIVs (um baseado na Distância de Bhattacharyya na forma de uma variável por vez e outro na Informação Mútua das variáveis) e de três ferramentas classificadoras (Máquina de Suporte Vetorial, K-Vizinhos Próximos e Distância de Mahalanobis) em uma eliminação *backward* ordenada pelos IIVs. Para obter resultados mais aprofundados das combinações deste artigo, as variações são testadas em dados reais e dados simulados com diferentes níveis de correlação e ruído.

1.2 Objetivos

O objetivo principal da dissertação consiste em propor métodos para seleção de variáveis com vistas à classificação de bateladas em categorias.

Os objetivos específicos são:

- Criar novos Índices de Importância de Variáveis com vistas à remoção ordenada de variáveis;
- Avaliar o desempenho de distintas ferramentas de classificação de observações;
- Comparar os resultados dos métodos a sistemáticas de seleção de variáveis mais difundidas;
- Avaliar a robustez dos métodos para seleção de variáveis propostos em bancos de dados reais e simulados caracterizados por distintos níveis de covariância e de ruído.

1.3 Justificativa do Estudo

Visando o aumento da qualidade do processo de produção, empresas investem cada vez mais em tecnologias de coleta e monitoramento de dados oriundos de processos. Tais dados, no entanto, acabam por gerar bancos em que ferramentas multivariadas de análise perdem eficiência por conta dos elevados níveis de correlação e ruído. Dentro deste contexto, o desenvolvimento de novas técnicas de seleção de variáveis se justifica pela necessidade de reduzir a dimensionalidade de dados, tornando viável a execução de análises multivariadas, sem a perda de informações relevantes. Além disso, bancos de dados reduzidos facilitam o monitoramento de processos, permitindo não apenas identificar previamente alterações de comportamento do processo, como também oferecendo condições de classificar bateladas corretamente de acordo com as especificações desejadas (KOURTI; MACGREGOR, 1995; MARTIN *et al.,* 1999; GAUCHI; CHAGNON, 2001; KETTANEH *et al.*, 2005; LIU; YU, 2005; ANZANELLO, 2009).

No contexto acadêmico, percebe-se um grande esforço devotado ao desenvolvimento de abordagens com vistas à seleção de variáveis em diversas áreas do conhecimento: Yang e Pedersen (1997) para categorização de texto, Rebolo *et al.* (2000) para categorização de vinhos, Westad *et al.* (2003) para avaliar a opinião de consumidores, Chen *et al.* (2011) para o diagnóstico de câncer de mama e Anzanello *et al.* (2013) para identificar falsificações de remédios. A possibilidade de aprimoramento das técnicas existentes através da combinação de novas técnicas e ferramentas voltadas ao contexto industrial justifica o desenvolvimento deste estudo no âmbito acadêmico.

1.4 Procedimentos Metodológicos

A presente pesquisa é classificada como aplicada, pois objetiva gerar conhecimentos para aplicação prática e dirigidos à solução de problemas específicos e de abordagem quantitativa (SILVA; MENEZES, 2005). O trabalho utiliza como procedimento o estudo de caso, uma vez que permite o amplo e detalhado conhecimento de um cenário prático (GIL, 2010).

1.5 Estrutura da Dissertação

A dissertação está organizada em cinco capítulos. O primeiro capítulo introduz o trabalho, apresentando objetivos, justificativas e o método de pesquisa adotado, sendo complementado pela delimitação do estudo e pela estrutura do trabalho.

O segundo capítulo apresenta o primeiro artigo, que propõe um novo método para seleção de variáveis com vistas à classificação de bateladas produtivas em duas categorias de qualidade. Para tanto, um Índice de Importância de Variáveis (IIV) é proposto com base nos parâmetros oriundos da Análise de Componentes Principais. Tal índice orienta a remoção das variáveis uma a uma, gerando um novo subconjunto a cada iteração. Estes subconjuntos são classificados pela Distância de Mahalanobis, sendo considerado como ótimo o subconjunto que estiver mais próximo de um hipotético resultado ideal. O método proposto é aplicado em cinco bancos de dados reais e seus resultados são comparados com a acurácia da ferramenta classificadora sem a remoção de variáveis e com a acurácia geradas pela sistemática *stepwise* de seleção de variáveis.

O terceiro capítulo traz o segundo artigo, que introduz uma revisão sobre os princípios teóricos da seleção de variáveis. O método proposto neste artigo é uma variação do método do primeiro artigo, com a substituição da medida estatística Distância de Mahalanobis pela técnica de aprendizagem computacional Máquina de Suporte Vetorial. O método proposto é aplicado em quatro bancos de dados reais e seus resultados são validados através da comparação com a acurácia da ferramenta classificadora sem a remoção de variáveis e com a acurácia da metodologia Omita Uma Variável por Vez executada com a mesma ferramenta classificadora.

O quarto capítulo apresenta o terceiro artigo, que compara variações da metodologia utilizada nos artigos anteriores utilizando dois novos IIVs e três ferramentas classificadoras. Para identificar o comportamento das variações, elas são aplicadas em cinco bancos de dados reais e bancos de dados simulados com diferentes níveis de covariância e ruído.

O quinto e último capítulo apresenta as conclusões do trabalho, avaliando os resultados obtidos frente aos objetivos e limitações do estudo, e sugestões para futuros desdobramentos desta pesquisa.

1.6 Delimitações do Estudo

Constituem restrições do presente estudo:

- A seleção de variáveis ocorre apenas por meio de modelos *wrapper* com eliminação *backward* ordenada;
- Os bancos de dados utilizados são aproximadamente balanceados;
- Não são feitas classificações em múltiplas categorias; e
- As variáveis são selecionadas com o objetivo de classificação e não de predição.

1.7 Referências Bibliográficas

ANZANELLO, M. J. Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes. **Gestão & Produção**, v.16, n.4, p.526-533, 2009.

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97-105, 2012.

ANZANELLO, M.J.; ORTIZ, R.S.; LIMBERGERB, R.P.; MAYORGA, P. A MULTIVARIATE-BASED Wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v.83, n.1, p.209-214, 2013.

BLUM, A.L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v.97, n.1, p.245-271, 1997.

BRODNJAK-VONCINA, D.; KODBA, Z.C.; NOVIC, M. Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. **Chemometrics and Intelligent Laboratory Systems**, v. 75, n. 1, p. 31-43, 2005.

CHEN, H.L.; YANG, B.; LIU, J.; LIU, D.Y. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v.38, n.7, p.9014-9022, 2011.

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v.58, n.2, p.171-193, 2001.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010. 200p.

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

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.; FRANKLIN, J. The elements of statistical learning: data mining, inference and prediction. **The Mathematical Intelligencer**, v.27, n.2, p.83-85, 2005.

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

KOURTI, T.; MACGREGOR, J.F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics and Intelligent Laboratory Systems**, v.28, n.1, p.3-21, 1995.

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

MARTIN, E.B.; MORRIS, A.J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, v.23, n.1, p.35-44, 1999.

PEREIRA, A.C.; REIS, M.S.; SARAIVA, P.M.; MARQUES, J.C. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. **Chemometrics and Intelligent Laboratory Systems**, v.105, n.1, p.43-55, 2011.

REBOLO, S.; PENA, R.; LATORRE, M.; BOTANA, A.; HERRERO, C. Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. **Analytica Chimica Acta**, v.417, n.2, p.211-220, 2000.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v.23, n.19, p.2507-2517, 2007.

SILVA, E. L.; MENEZES, E. M. Metodologia da Pesquisa e Elaboração de Dissertação. 3. ed. Florianópolis: Laboratório de Ensino à Distância da UFSC, 2001. 121 p.

TIAN, W.M.; HE, Z.; YAN, W. Key Process Variable Identification for Quality Classification Based on PLSR Model and Wrapper Feature Selection. **In Proceedings of 2012 3rd International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012)**, p.263-270. Springer Berlin Heidelberg, 2013.

WESTAD, F.; HERSLETH, M.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v. 14, n. 5, p. 463-472, 2003.

YANG, Y. AND PEDERSEN, J.O. 1997. A comparative study on feature selection in text categorization. **In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN)**, 412-420, 1997.

2. Primeiro artigo: Sistemática de seleção de variáveis para classificação de bateladas produtivas

Alessandro Kahmann

Michel José Anzanello

Universidade Federal do Rio Grande do Sul

Resumo

A seleção de variáveis é um importante problema a ser analisado no monitoramento de processos industriais. Tais processos tipicamente produzem bancos com dados ruidosos ou altamente correlacionados, prejudicando sua análise. Em processos que operam em bateladas, esta necessidade se acentua por conta de cenários onde o número de variáveis é maior que o de observações, comprometendo a eficiência de diversos métodos multivariados. Neste artigo é proposto um novo método para seleção de variáveis (Remoção Ordenada para seleção de Variáveis - ROV) com vistas à classificação de bateladas produtivas em duas categorias de qualidade. Para tanto, um Índice de Importância de Variáveis (IIV) é proposto com base nos parâmetros oriundos da Análise de Componentes Principais (ACP); tal índice servirá para o ordenamento das variáveis de processo de acordo com sua presumida capacidade de discriminar bateladas. Na sequência, as variáveis são excluídas uma a uma de acordo com a ordem sugerida pelo IIV; a cada eliminação, uma nova classificação é realizada através da Distância de Mahalanobis e a acurácia de classificação é reavaliada. O procedimento é interrompido quando houver apenas uma variável remanescente. Ao ser aplicado em cinco bancos de dados industriais, a sistemática proposta resultou em um incremento médio de 28,4% na acurácia em relação à classificação com todas as variáveis, com a retenção média de 8,24% das variáveis originais.

Palavras-chave: Análise de Componentes Principais, Distância de Mahalanobis, Seleção de variáveis.

Abstract

Variable selection is an important issue to be analyzed in industrial processes monitoring. Such processes typically produce databases with noisy or highly correlated data, jeopardizing their analysis. In processes that operate in batches, this need is accentuated due to scenarios where the number of variables is larger than the number of samples, reducing the efficiency of several multivariate methods. This paper introduces a new method for variable selection (ROV) in order to classify production batches into two quality classes. For that matter, a Variable Importance Index (VII) is generated based on Principal Components Analysis (PCA) weights; this index ranks process variables according to their presumed ability to discriminate batches. Further, variables are removed one by one according to the order suggested by VII; after each variable removal, a new classification is realized through the Mahalanobis Distance, and the classification accuracy is reevaluated. The iterative procedure is repeated until there is only one variable left. When applied to five industrial datasets, the proposed systematic increases average accuracy 28,4% compared with the classification with all variables, and retained average 8,24% of original variables.

Keywords: Principal Components Analysis, Mahalanobis Distance, Variable Selection.

2.1 Introdução

Com o rápido avanço de tecnologias para geração, monitoramento e análise de bancos de dados, são obtidas cada vez mais informações que viabilizam a identificação de padrões que expliquem eventos das mais diversas naturezas. Segundo Liu e Yu (2005), este avanço tipicamente gera bancos de dados com elevado número de variáveis (desencorajando uma análise minuciosa das mesmas), ou conduz a situações onde ferramentas de análise perdem eficiência frente a dados impregnados por ruído ou altamente correlacionados. Dentro deste contexto, a utilização de métodos para seleção de variáveis visa reduzir tais dados a uma quantidade adequada, que permita executar análises sem a perda de informações importantes. Em aplicações industriais, tal resultado é almejado para a viabilidade de monitoramento e controle de processos.

Anzanello *et al*. (2009; 2012) citam que, em processos industriais, pode haver centenas ou até milhares de variáveis ruidosas ou correlatas, incluindo temperatura, pressão,

concentração de componentes e tempos de reação, entre outros. O grande volume de dados coletados de tais processos desafia pesquisadores a desenvolverem abordagens eficientes para avaliar tais processos, justificando o grande número de estudos devotados a selecionar as variáveis que melhor explicam determinados processos em áreas da engenharias (GAUCHI; CHAGNON, 2001; WESTAD *et al*., 2003; URTUBIA *et al*., 2007), saúde (BLOCK *et al*., 1998; AKAY, 2009), e estatística (RUGGIERI; LAWRENCE, 2012; HAPFELMEIER; ULM, 2013), entre outras.

Sistemáticas de seleção de variáveis com aplicação em cenários industriais podem ser separadas em duas frentes: (*i*) seleção de variáveis para a predição, na qual o objetivo é encontrar um conjunto limitado de variáveis independentes *x'*s, que viabilizam melhor predição de uma ou mais variáveis dependentes *y'*s, como em Wold *et al*. (2001), Gauchi e Chagnon (2001) e Pereira *et al*. (2011); e (*ii*) seleção de variáveis para classificação, na qual o objetivo é encontrar o conjunto de variáveis independentes com a melhor habilidade discriminante para categorizar observações entre classes, como em Brodnjak-Voncina *et al*. (2005), Urtubia *et al.* (2007) e Anzanello *et al*. (2009; 2012). As proposições deste artigo estão alinhadas com a segunda frente.

Neste artigo é apresentado um método para seleção de variáveis com propósito de classificação de bateladas produtivas em duas categorias, operacionalizado através dos seguintes passos: (1) separar o banco de dados em porções de treino e teste; (2) na porção de treino, aplicar a ACP para a extração de informações a respeito da influência das variáveis na explicação da variabilidade do processo; os parâmetros da ACP geram um índice de importância das variáveis; (3) aplicar a distância de Mahalanobis para a classificação dos dados em dois grupos utilizando todas as variáveis. Após classificadas as bateladas e calculada a acurácia das classificações, remover a variável com o menor IIV e repetir a classificação com as variáveis remanescentes; repetir esse procedimento iterativo até restar apenas uma variável; e (4) definir o conjunto de variáveis classificatórias responsáveis pela maior acurácia e menor percentual de variáveis retidas. Utilizando as variáveis selecionadas, classificar o banco de teste.

Este artigo está estruturado como segue. Na seção seguinte são apresentadas as ferramentas utilizadas na metodologia proposta. Na terceira seção é detalhado o método proposto, explicando sua operacionalização. Na quarta seção são apresentados os resultados

da aplicação do método em dados industriais. Na quinta seção são mostradas as conclusões oriundas deste estudo e propostas para trabalhos futuros.

2.2 Fundamentação Teórica

Nesta seção são apresentados os fundamentos teóricos da ACP e da Distância de Mahalanobis, ferramentas estatísticas utilizadas no método de seleção de variáveis proposto neste artigo.

2.2.1 Análise de Componentes Principais (ACP)

A ACP é uma ferramenta multivariada que permite redução de dados e extração de características de bancos de dados caracterizados por elevados níveis de ruído e dados espúrios (YANG; WANG, 1999; ESBENSEN, 2002; DHARMARAJ *et al*., 2006).

A ACP tem como objetivo substituir um conjunto de *N* variáveis correlacionadas, $x_1, x_2, \dots, x_N$, por um conjunto de variáveis não correlacionadas, $t_1, t_2, \dots, t_N$, chamados de componentes principais. Estas novas variáveis, cuja estrutura de correlação é mais simples, são combinações lineares das variáveis originais, arranjadas de tal modo que suas variâncias estejam em ordem decrescente de grandeza e a variância total do conjunto inicial das variáveis seja preservada (NETO; MOITA, 1998; GOMES *et al.*, 2004; DHARMARAJ *et al*., 2006). A representação genérica da combinação linear é trazida na equação (1),

$$t_i = \sum_{j=1}^{N} w_{ij}\, x_j, \quad \text{onde } i=1,2,\dots,N \tag{1}$$

onde w_{ij} são os pesos da variável x_j em determinada combinação linear. Cada componente t_i possui uma variância λ_i associada. Da soma destas variâncias pode-se encontrar a porcentagem de variância explicada de cada componente. A soma acumulada dos valores das variâncias de cada componente principal podem ser plotados em um gráfico que permite visualizar agrupamentos ou tendências no banco de dados, o que auxilia na decisão de quantos componentes serão retidos (MASSART *et al.*, 1998).

Dentro dos *N* componentes são verdadeiras as propriedades:

a) $\text{Cov}\,(t_i, t_l) = 0, \forall\, i,l=1,2,\dots N : i \neq l$
b) $\text{Var}\,(t_1) \geq \text{Var}\,(t_2) \geq \cdots \geq \text{Var}\,(t_N)$
c) $\sum_{i=1}^{N} \text{Var}\,(t_i) = \sum_{l=1}^{N} \text{Var}\,(x_l)$

Apesar desta técnica gerar *N* componentes, tipicamente os dois ou três primeiros componentes explicam grande porcentagem da variância total original. Em Câmara *et al.* (2006), os dois primeiros componentes explicam 81,2% da variabilidade original; enquanto em Brodnjak-Voncina *et al*. (2005), a retenção dos dois primeiros componentes mantém 97,8% da variabilidade. Já a retenção dos três primeiros componentes, em Urtubia *et al.* (2007), mantém 70% da variabilidade; em Nataraja e Johnson (2011), 80%; em Westad *et al.* (2003), 91,6% e em Diaz *et al*. (2005), 98,9%.

2.2.2 Distância de Mahalanobis

A distância Euclidiana é base para ferramentas classificadoras, como *Kmeans* e *K-nearest neighbor* (KNN), que utilizam esta medida para definir sua classificação. A distância Euclidiana assume, no entanto, que cada variável é igualmente importante e independente, o que nem sempre ocorre em casos reais. Por estes motivos, esta distância pode não gerar resultados acurados quando aplicada em casos multivariados (XIANG *et al*., 2008).

A distância de Mahalanobis baseia-se na distância Euclidiana, porém pondera os elementos da amostra pela matriz de covariância. Esta medida forma elipsoides ao redor do centroide, determinando escalas de distâncias da observação analisada até os respectivos centroides de cada grupo, como ilustrado na Figura 2.1. Para o caso de classificação, uma nova observação é inserida na classe cuja distância de Mahalanobis for menor (MAESSCHALCK *et al*., 2000; DIXON; BRERETON, 2009).

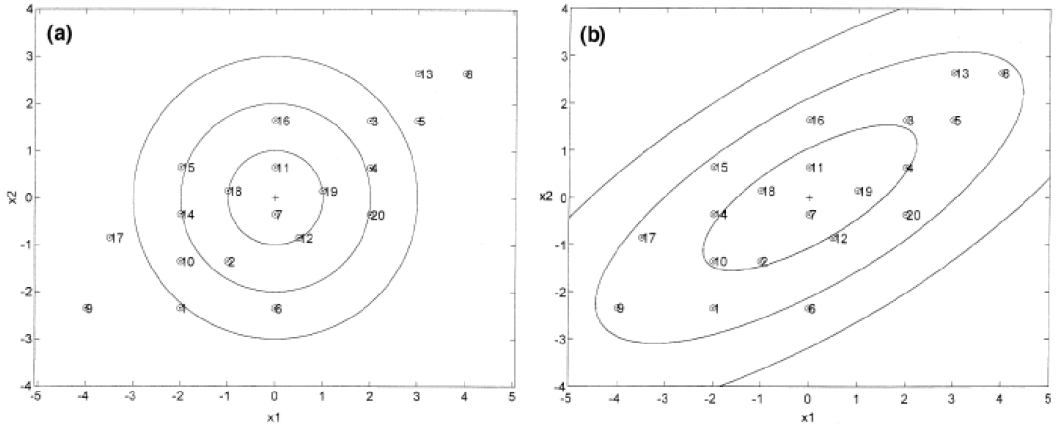

Figura 2.1 - (a) Distância de dados ao centro de acordo com a Distância Euclidiana; (b) Distância de dados ao centro de acordo com a Distância de Mahalanobis
Fonte: Maesschalck *et al*., 2000

O cálculo da distância de Mahalanobis entre uma observação e um grupo de observações é realizado através da equação (2)

$$dm^2_{(amostra,grupo)} = (\boldsymbol{x} - \mu)^T \sum^{-1} (\boldsymbol{x} - \mu) \quad (2)$$

onde **x** é o vetor de variáveis da observação a ser classificada, μ é o vetor média das variáveis da matriz do grupo em comparação e $\sum$ é a matriz de covariância da matriz dos elementos do grupo em comparação. Há casos em que a matriz de covariância é singular e, consequentemente, não invertível. Para estes casos, Bohling *et al*. (1998) utiliza a matriz pseudo-inversa de Moore-Penrose, mesmo recurso utilizado em diversos *softwares*.

2.3 Metodologia para seleção de variáveis

O método de seleção de variáveis para classificação proposto neste artigo, chamado de Remoção Ordenada para Seleção de Variáveis (ROV), consiste em quatro passos, detalhados na sequência.

Passo 1: Separar os dados históricos em bancos de treino e teste

Considere a matriz $X_{(M,N)}$, onde *M* representa o número de observações (bateladas) e *N* o número de variáveis de processo (independentes). A partir desta matriz, são randomicamente criadas duas matrizes *Tr* e *Ts*, onde *Tr* ∪ *Ts* = *X*. A porção de treino (*Tr*) é utilizada para encontrar o subconjunto de variáveis que melhor classifica o processo, enquanto a de teste (*Ts*) é utilizada para a verificação de acurácia do método. A quantidade de observações em cada matriz é definida pelas proporções 90-10, 80-20, 70-30, 60-40 e 50-50; diferentes proporções de treino e teste buscam avaliar a influência deste fator no método de seleção proposto.

Passo 2: Gerar o IIV

O IIV é utilizado para ordenar a remoção de variáveis não importantes ou que prejudiquem a classificação. O principal objetivo de ranquear as variáveis é prevenir correlações espúrias que podem ser tratadas como informações relevantes (WESTAD *et al*., 2003).

O IIV proposto baseia-se nos pesos w_{ij} das *N* variáveis nos *p* componentes principais retidos. Cada w_{ij} é ponderado pela variância λ_i associada ao componente. Como a análise é

baseada na influência, tanto positiva quanto negativa, de w_{ij} nos p componentes retidos, estes valores são somados na forma absoluta. O IIV associado a cada variável j é gerado pela equação (3).

$$IIV_j = \sum_{i=1}^{p} |w|_{ij} * \lambda_i \quad (3)$$

Assume-se que uma variável com menor IIV explique menos variabilidade do processo, ou é redundante a outra variável com maior IIV, o que pode resultar em uma análise menos precisa.

Em Anzanello *et al*. (2011), tem-se um IIV semelhante, em que o índice é constituído apenas pela soma absoluta dos pesos. Verifica-se, no entanto, que componentes principais que pouco explicam a variabilidade do processo podem ter pesos altos em determinadas variáveis. Desta forma, a ponderação por λ_i diminui esta influência em casos em que é necessário reter mais de um componente principal para obter uma explicação de variabilidade aceitável.

Passo 3: Classificação das amostras do banco de treino em duas classes, e remoção de variáveis irrelevantes

Nesta etapa as bateladas são classificadas como "conforme" e "não conforme" através da distância de Mahalanobis. Inicialmente, as amostras pertencentes à matriz *Tr* classificadas como "conforme" são alocadas na matriz *C*, enquanto as classificadas com "não conforme" são alocadas na matriz *NC*. As matrizes *C* e *NC* são a base para a classificação das amostras, tanto da porção de treino quanto da porção de teste.

Na sequência, são calculadas as distâncias dm^2 entre as amostras da matriz *Tr* até as matrizes *C* e *NC*. A amostra será considerada como pertencente à classe a qual sua distância for menor. Após classificadas todas as amostras da matriz *Tr*, remove-se a variável com menor IIV das matrizes *C, NC* e Tr e classificam-se novamente as observações. As variáveis são removidas até restar apenas uma.

A equação (4) mostra o cálculo da acurácia do método a cada iteração.

$$\text{ACC} = \frac{\text{Classificações corretas}}{\text{Total de amostras}} \quad (4)$$

Passo 4: Determinar o melhor subconjunto de variáveis e classificar o banco de teste com as variáveis retidas

Para selecionar o melhor conjunto de variáveis é utilizada uma simplificação da ferramenta existente em Anzanello *et al*. (2012). A seleção consiste em encontrar o ponto, de coordenadas (% variáveis retidas, acurácia), com a menor distância euclidiana a um ponto ótimo hipotético definido pelo usuário como ideal. Como ambas as medidas de desempenho, acurácia e retenção de variáveis, estão no intervalo [0,1], é considerado ótimo um cenário onde apenas uma variável retida resulta em uma acurácia de 100% (logo, o ponto ótimo tem como coordenadas (1/(número de variáveis),1). O subconjunto de variáveis mais próximo ao ponto ótimo forma o subconjunto recomendado de variáveis, como ilustrado na Figura 2.2.

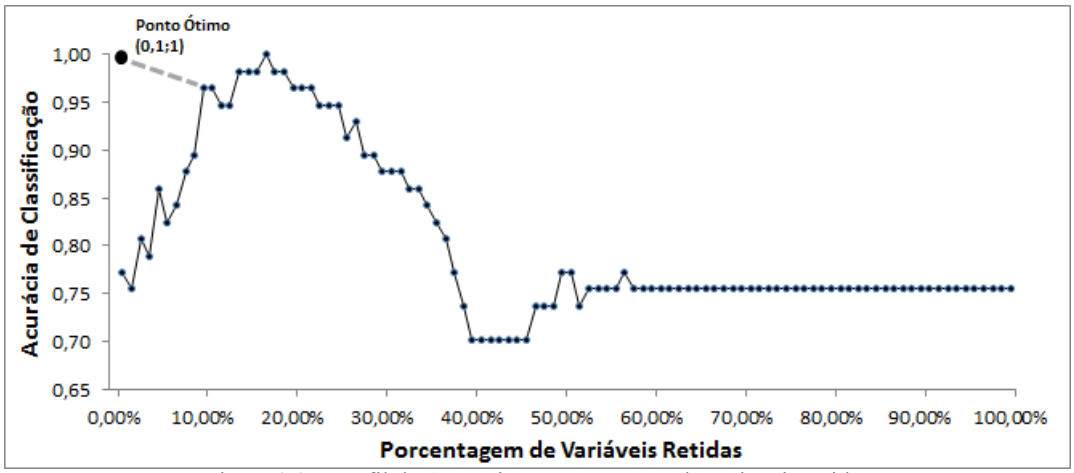

Figura 2.2 – Perfil de Acurácia x Porcentagem de variáveis retidas
Fonte: Os autores

Definido o subconjunto recomendado de variáveis, classificam-se as amostras do banco de teste, que representam novas bateladas não incluídas na geração do modelo. Avalia-se a acurácia do banco de teste por meio da equação (4).

2.4 Resultados

O método proposto é aplicado em cinco bancos de dados, disponíveis em Gauchi e Chagnon (2001) e apresentados na Tabela 2.1. Tais bancos são compostos por *M* bateladas industriais descritas por *N* variáveis de processo, sendo referidos como ADPN, LATEX, OXY, SPIRA e GRANU.

O banco de dados ADPN advém da produção industrial de adiponitrila; LATEX descreve um estágio da polimerização da produção de látex; OXY foi obtido de um processo

de obtenção de dióxido de titânio; SPIRA apresenta dados de um processo de fermentação utilizado na produção de antibióticos, e GRANU foi retirado de um processo de fabricação de emulsões antiespumantes utilizadas na indústria do papel. Mais detalhes sobre os bancos de dados são encontrados em Gauchi e Chagnon (2001).

Tabela 2.1 - Bancos de dados testados

Banco de dados	Número de observações	Número de variáveis no processo
ADPN	71	100
LATEX	262	117
OXY	25	95
SPIRA	145	96
GRANU	29	78

Com vistas a uma melhor avaliação de desempenho do método, os dados históricos foram aleatoriamente permutados entre os bancos de treino e teste, mantendo-se o tamanho de cada banco, porém com diferentes amostras a cada análise, em 300 repetições para cada banco de dados. Objetiva-se não favorecer ou prejudicar o método com uma única separação dos bancos em treino e teste. A análise foi realizada no Matlab® versão 7.8.0 e os códigos foram criados pelos autores.

O total de variância retida pela ACP variou entre 80% e 95% de acordo com o banco de dados em estudo. A Tabela 2.2 mostra a acurácia média da ferramenta de classificação (distância de Mahalanobis) aplicada a todas as variáveis originais de cada banco, enquanto a Tabela 2.3 mostra a acurácia média e percentual médio de variáveis retidas (entre parêntesis) de dois métodos de seleção: a Remoção Ordenada de Variáveis (ROV), proposta neste artigo, e a regressão *Stepwise* (SW), técnica popularmente utilizada para seleção de variáveis que retém as variáveis com determinado nível de significância estatística, a um nível de 5%. Ambas as tabelas apresentam os resultados para distintas proporções de treino e teste.

Tabela 2.2 - Acurácia com todas as variáveis

Banco de dados	Acurácia média na proporção treino-teste				
	90-10	80-20	70-30	60-40	50-50
ADPN	0,40	0,40	0,44	0,49	0,50
LATEX	0,79	0,78	0,78	0,76	0,74
OXY	0,29	0,27	0,30	0,30	0,31
SPIRA	0,38	0,40	0,40	0,40	0,37
GRANU	0,62	0,56	0,56	0,55	0,56
MÉDIA	0,49	0,48	0,49	0,50	0,50

Tabela 2.3 - Acurácia média e (porcentual médio de variáveis retidas) dos métodos

Banco de dados	Acurácia média e porcentual médio de variáveis retidas na proporção treino-teste									
	90-10		80-20		70-30		60-40		50-50	
	ROV	SW	ROV	SW	ROV	SW	ROV	SW	ROV	SW
ADPN	0,78 (8,31)	0,79 (17,65)	0,78 (8,28)	0,79 (16,69)	0,78 (8,28)	0,79 (15,15)	0,79 (7,95)	0,79 (14,05)	0,79 (7,68)	0,78 (13,61)
LATEX	0,80 (11,09)	0,83 (17,12)	0,80 (11,58)	0,83 (16,66)	0,80 (11,42)	0,82 (15,33)	0,79 (11,16)	0,82 (13,57)	0,80 (11,14)	0,81 (11,98)
OXY	0,80 (4,22)	0,81 (4,37)	0,82 (4,18)	0,81 (4,67)	0,85 (3,63)	0,82 (4,40)	0,85 (3,38)	0,82 (4,83)	0,84 (3,09)	0,82 (4,71)
SPIRA	0,73 (10,48)	0,72 (7,64)	0,72 (10,61)	0,72 (7,45)	0,72 (11,00)	0,73 (7,47)	0,72 (10,85)	0,73 (7,22)	0,72 (10,62)	0,73 (7,38)
GRANU	0,77 (8,14)	0,68 (3,20)	0,74 (7,96)	0,70 (2,82)	0,75 (7,68)	0,70 (2,60)	0,72 (7,02)	0,70 (2,64)	0,74 (6,23)	0,69 (2,64)
MÉDIA	0,78 (8,45)	0,77 (9,99)	0,77 (8,52)	0,77 (9,66)	0,78 (8,40)	0,77 (8,99)	0,77 (8,07)	0,77 (8,46)	0,78 (7,75)	0,76 (8,06)

Da comparação da Tabela 2.2 com a Tabela 2.3, verifica-se ganho expressivo com a utilização dos métodos de seleção de variáveis em todos os bancos de dados, quando comparado com a aplicação da distância de Mahalanobis sem a remoção de variáveis. O incremento médio de acurácia é de 28,4%, com a retenção média de 8,24% das variáveis originais através do ROV; tais valores são superiores ao método SW, o qual eleva a acurácia média em 27,6%, com a retenção média de 9,30% das variáveis originais.

Da Tabela 2.3 nota-se ainda que não há diferenças geradas pelas distintas proporções de treino e teste em termos de acurácia e retenção de variáveis. Isso é justificado pelo fato da distância de Mahalanobis ser baseada no centroide do conjunto, que é definido majoritariamente pelos elementos mais próximos a ele (DIXON; BRERETON, 2009). Logo, mesmo possuindo menor número de observações na porção de treino, o centroide dos conjuntos “conforme” e “não conforme” é definido por poucas destas observações, variando pouco sua posição. Portanto este método é aplicável, sem perda de acurácia, em casos com poucas observações, característica de processos que operam em bateladas.

O subconjunto de variáveis recomendado varia a cada repetição tanto no valor de acurácia quanto na quantidade de variáveis retidas, por conta da aleatorização das observações inseridas nos conjuntos de treino e teste. Para melhor avaliar o método, na Tabela 2.4 e Tabela 2.5 são analisadas as variabilidades (desvio-padrão) da acurácia e percentual de variáveis retidas nas 300 repetições executadas.

Tabela 2.4 - Desvio padrão da acurácia

Banco de dados	Desvio padrão da acurácia									
	90-10		80-20		70-30		60-40		50-50	
	ROV	SW	ROV	SW	ROV	SW	ROV	SW	ROV	SW
ADPN	0,08	0,13	0,07	0,10	0,06	0,08	0,04	0,07	0,04	0,07
LATEX	0,05	0,06	0,04	0,04	0,04	0,03	0,03	0,03	0,03	0,03
OXY	0,11	0,23	0,09	0,16	0,08	0,14	0,07	0,13	0,08	0,13
SPIRA	0,07	0,10	0,06	0,07	0,04	0,06	0,04	0,06	0,03	0,05
GRANU	0,13	0,26	0,10	0,18	0,09	0,15	0,07	0,13	0,06	0,12
MÉDIA	0,09	0,16	0,07	0,11	0,06	0,09	0,05	0,08	0,05	0,08

Tabela 2.5 - Desvio padrão da porcentagem de variáveis retidas

Banco de dados	Desvio padrão das variáveis retidas									
	90-10		80-20		70-30		60-40		50-50	
	ROV	SW	ROV	SW	ROV	SW	ROV	SW	ROV	SW
ADPN	1,24	4,18	1,51	5,39	1,62	5,43	1,69	5,75	1,90	6,29
LATEX	2,14	2,29	2,21	2,50	2,00	2,74	2,09	2,67	1,87	2,72
OXY	0,97	2,80	1,05	2,27	1,18	2,44	1,21	3,06	1,21	3,02
SPIRA	1,77	2,09	1,68	2,11	1,93	2,22	2,04	1,92	2,23	2,60
GRANU	1,39	1,09	2,03	1,24	2,09	1,27	2,37	1,58	2,30	1,74
MÉDIA	1,50	2,49	1,70	2,70	1,76	2,82	1,88	2,99	1,90	3,27

Utilizando os níveis definidos em Gomes (1985), os coeficientes de variação das acurácias são classificados como baixo (abaixo de 0,10) ou normal (entre 0,10 e 0,20), o que permite concluir que o método proposto produz resultados pouco dispersos em torno da média. Isso mostra a solidez do método, uma vez que o mesmo não produz resultados discrepantes em relação à média, o que pode prejudicar o controle do processo. Já os coeficientes de variação da porcentagem de variáveis retidas são classificados como médios ou altos (entre 0,20 e 0,30), porém tal análise é comprometida pelo fato da média ser do tipo menor é melhor e baixa em relação à própria escala.

Em comparação com o método SW, o ROV produz resultados menos dispersos em torno da média, tanto na acurácia média quanto na porcentagem média de retenção de variáveis. Pode-se então afirmar que o ROV produz resultados melhores e mais consistentes.

2.5 Conclusão

Foi proposto neste artigo um novo método para selecionar as variáveis que melhor classificam bateladas produtivas descritas por variáveis correlatas e ruidosas, intitulado Remoção Ordenada de Variáveis (ROV). Este método destaca-se pela simplicidade teórica das ferramentas estatísticas utilizadas e consiste em (1) separar os dados históricos em bancos de treino e teste; (2) aplicar a ACP nos dados e gerar o IIV através dos pesos relacionados às variáveis dos componentes retidos, ponderados pela sua variância explicada; (3) classificar as

amostras através da distância de Mahalanobis, calcular a acurácia da classificação, retirar a variável com menor IIV e repetir a classificação até restar apenas uma variável; (4) definir o subconjunto de variáveis que melhor classifica o processo e verificar a acurácia do método utilizando as variáveis selecionadas para classificar o banco de teste.

Ao ser aplicado em cinco processos industriais, o método obteve uma melhora média de 28,4% na classificação das observações, com a retenção de 8,24% das variáveis. O método mostrou-se eficiente nas diversas proporções de tamanhos dos bancos de treino e teste avaliadas, mostrando que pode ser aplicado em situações caracterizadas pela presença de poucas observações. Quando comparado com o método *Stepwise*, sistemática amplamente difundida para seleção de variáveis, o método ROV produziu resultados superiores e mais confiáveis.

Possibilidades de pesquisas futuras englobam a substituição da ferramenta classificadora para comparação de eficácia da remoção de variáveis com ferramentas de características diferentes. Já uma continuação desta pesquisa inclui a extensão do método a casos de múltiplas classes, onde as bateladas podem ser inseridas em grupos "não conforme", "regular" e "premium", uma vez que distância de Mahalanobis é aplicável a tal situação.

2.6 Referências

AKAY, M.F. Support vector machines with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, n. 2, p. 3240-3247, 2009.

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97-105, 2012.

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. 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.

ANZANELLO, M.J.; FOGLIATTO, F.S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preferences**, v. 22, n. 1, p. 139-148, 2011.

BLOCK, P.C.; PETERSON, E.C.; KRONE, R.; KESLER, K.; HANNAN, E.; O'CONNOR, G.T.; DETRE, KATHERINE. Identification of variables needed to risk adjust outcomes of coronary interventions: Evidence-based guidelines for efficient data collection. **Journal of the American College of Cardiology**, n. 32, v. 1, p. 275-282, 1998.

BOHLING, G.C.; DAVIS, J.C.; OLEA, R.A.; HARFF, J. Singularity and Nonnormality in the Classification of Compositional Data. **Mathematical Geology**, v. 30, n. 1, p. 5-20, 1998.

BRODNJAK-VONCINA, D.; KODBA, Z.C.; NOVIC, M. Multivariate data analysis in classification of vegetable oils characterized by the content of fatty acids. **Chemometrics and Intelligent Laboratory Systems**, v. 75, n. 1, p. 31-43, 2005.

CÂMARA, J.S.; ALVES, M.A.; MARQUES, J.C. Multivariate analysis for the classification and differentiation of Madeira wines according to main grape varieties. **Talanta**, v. 68, n. 5, p. 1512-1521, 2006.

DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D.L. The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems**, v. 50, n. 1, p. 1-18, 2000.

DIAZ, T.G.; MERÁS, I.D.; CASAS, J. S.; FRANCO, M.F.A. Characterization of virgin olive oils according to its triglycerides and sterols composition by chemometric methods. **Food Control**, v. 16, n. 4, p. 339-347, 2005.

DIXON, S.; BRERETON, R.G. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. **Chemometrics and Intelligent Laboratory Systems**, v. 95, n. 1, p. 1-17, 2009.

DHARMARAJ, S.; HOSSAIN, M. A.; ZHARI, S.; HARN, G. L.; ISMAIL, Z. The use of principal component analysis and self-organizing map to monitor inhibition of calcium oxalate crystal growth by Orthosiphon stamineus extract. **Chemometrics and Intelligent Laboratory Systems**, v. 81, n. 1, p. 21-28, 2006.

ESBENSEN, K.H. **Multivariate Data Analysis - In Practice**. 5ª ed. Oslo: CAMO Process AS, 2002.

HAPFELMEIER, A.; ULM, K. A new variable selection approach using Random Forest. **Computational Statistics and Data Analysis**, v. 60, p. 50-69, 2013.

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 171-193, 2001.

GOMES, F.P. **Curso de estatística experimental**.11ª ed. Piracicaba: Livraria Nobel S.A., 1985.

GOMES, J.B.V.; CURI, N.; MOTTA, P.E.F.; KER, J.C.; MARQUES, J.J.G.S.M.; SCHULZE, D. G. Análise de componentes principais de atributos físicos, químicos e mineralógicos de solos do bioma cerrado. **Revista Brasileira de Ciências do Solo**, v. 28, n. 1, p. 137-153, 2004.

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

MASSART, D.L.; VANDEGINSTE, B.G.M.; DEMING, S.N.; MICHOTTE, Y.; KAUFMAN, L. **Chemometrics: A Textbook**. Amsterdam: Elsevier, 1988.

NATARAJA, N. R.; JOHNSON, A. L. Guidelines for using variable selection techniques in data envelopment analysis. **European Journal of Operational Research**, v. 215, n. 3, p. 662-669, 2011.

NETO, J.M.M.; MOITA, G.C. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, v.21, n.4, p. 467-469, 1998.

PEREIRA, A.C.; REIS, M.S.; SARAIVA, P.M.; MARQUES, J.C. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. **Chemometrics and Intelligent Laboratory Systems**, v. 105, n. 1, p. 43-55, 2011.

RUGGIERI, E.; LAWRENCE, C.E. On efficient calculations for Bayesian variable selection. **Computational Statistics and Data Analysis**, v. 56, n. 6, p. 1319-1332, 2012.

URTUBIA, A.; PÉREZ-CORREA, J.R.; SOTO, A.; PSZCZOLKOWSKI, P. Using data mining techniques to predict industrial wine problem fermentation. **Food Control**, v. 18, n. 12, p. 1512-1517, 2007.

WESTAD, F.; HERSLETH, M.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v. 14, n. 5, p. 463-472, 2003.

WOLD, S.; SJOSTROM, M.; ERIKSEN, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109-130, 2001.

XIANG, S.; NIE, F.; ZHANG, C. Learning a Mahalanobis distance metric for data clustering and classification. **Pattern Recognition**, v. 41, n. 12, p. 3600-3612, 2008.

YANG, T.; WANG, S. Robust algorithms for principal component analysis. **Pattern Recognition Letters**, v. 20, n. 9, p. 927-933, 1999.

3. Segundo artigo: Seleção de variáveis através de remoção ordenada para classificação de bateladas produtivas

Alessandro Kahmann

Michel José Anzanello

Universidade Federal do Rio Grande do Sul

Resumo

Um grande número de variáveis é normalmente usado para descrever processos industriais em diversas áreas. Tais dados tipicamente estão impregnados por variáveis ruidosas ou altamente correlacionadas, reduzindo a eficiência de diversas ferramentas multivariadas voltadas ao controle e monitoramento de processos. Em processos que operam em bateladas, esta dificuldade se acentua por conta de cenários onde o número de observações é menor que o número de variáveis. Para o estudo de tais bancos de dados, são utilizadas técnicas de mineração de dados associadas a sistemáticas de seleção de variáveis. Neste artigo é proposto um novo método de seleção de variáveis (Remoção Ordenada por Importância – ROI) com vistas à classificação de bateladas produtivas em duas categorias de qualidade. Para tanto, é proposto um Índice de Importância de Variáveis (IIV) com base nos parâmetros oriundos da Análise de Componentes Principais (ACP); tal índice ordenará as variáveis de acordo com sua presumida capacidade discriminante. Separadas em bancos de treino e teste, as variáveis do banco de treino são excluídas sistematicamente a cada iteração após classificação via Máquina de Suporte Vetorial (MSV). O procedimento é interrompido quando houver apenas uma variável remanescente. Um método avaliador, baseado na distância dos resultados dos subconjuntos a um hipotético resultado ideal, é utilizado para definir o subconjunto das variáveis originais que melhor classifica as amostras. Ao ser aplicado em quatro bancos de dados industriais, a sistemática reduziu o número de variáveis do processo em 92,53% e elevou a acurácia média de classificação em 25,14%, quando comparada à aplicação da ferramenta de classificação sem a exclusão de variáveis, e obteve melhores resultados em comparação à técnica “Omita Uma Variável por Vez” (OUVV).

Palavras-chave: Seleção de variáveis, Máquina de Suporte Vetorial, Análise de Componentes Principais, Classificação de bateladas produtivas

Abstract

A large number of variables is normally used to describe industrial process in many areas. Such data are typically impregnated with noise or highly correlated variables, reducing the efficiency of several multivariate tools tailored to process control and monitoring. In processes that operate in batch system, that limitation becomes crucial because the number of samples is typically lower than the number of variables. In order to study such datasets, data mining techniques are integrated to variable selection techniques. This paper proposes a new method of variables selection (ROI) to classify production batches into two quality levels. We propose a Variable Importance Index (VII) based on the parameters of Principal Component Analysis (PCA); such index ranks variables according to their presumed ability to discriminate batches. Separated in training ant testing sets, the variables of training set are removed systematically, and classified by the Support Vector Machine (SVM) technique. The procedure is stopped when there is only one remaining variable. We use a metric based on the distance of the results of the subset to a hypothetical optimum to define the recommended subset of original variables. When applied to four industrial datasets, the systematic reduced the number of process variables in 92,53% and increased the average accuracy in 25,14%, when compared to the classification tool with no variable selection. The proposed method also outperformed the “remove one variable at a time” technique.

Keywords: Variable Selection, Support Vector Machine, Principal Components Analysis, Production Batches Classification

3.1 Introdução

Mineração de dados, um subcampo interdisciplinar da ciência da computação, é o processo computacional para reconhecimento de padrões em grandes bancos de dados. Para tal processo, são utilizados métodos de aprendizagem computacional, ferramentas estatísticas e sistemas de bancos de dados. Com a evolução de computadores e tecnologias para armazenamento de dados, verifica-se a geração de bancos de dados com centenas, ou até

milhares, de variáveis ruidosas. Em determinadas aplicações práticas, como processos que operam com bateladas produtivas, o número de observações é tipicamente menor que o de variáveis, aumentando a dificuldade de análise destes dados. A mineração de dados tem por objetivo encontrar padrões e extrair informações deste vasto volume de dados (GUYON; ELISSEEFF, 2003; HASTIE *et al*., 2005; KETTANEH *et al.*, 2005; LIU; YU, 2005).

Segundo Blum e Langley (1997), a seleção de variáveis é uma das mais importantes, e frequentemente utilizadas, técnicas de mineração de dados. Ela tem por objetivo criar um modelo de análise, selecionando apenas variáveis significativas para o modelo, reduzindo a quantidade de atributos e, consequentemente, removendo dados ruidosos, redundantes, ou irrelevantes. Os efeitos desta redução são observados na facilitação da observação e compreensão dos dados e na maior rapidez e aumento de acurácia de algoritmos de mineração de dados (GUYON; ELISSEEFF, 2003; LIU; YU, 2005; HAPFELMEIER; CHEN *et al.*, 2011; ULM, 2013). A seleção de variáveis tornou-se objeto de estudo de várias áreas, como engenharias (GAUCHI; CHAGNON, 2001; LUTS *et al.,* 2003; WESTAD *et al*., 2003; ANZANELLO, 2009; ANZANELLO *et al.*, 2012), saúde (GREENLAND, 1989; AKAY, 2009; EVANS *et al.* 2012), estatística (ZOU; HASTIE, 2005; HAPFELMEIER; RUGGIERI; LAWRENCE, 2012; ULM, 2013) e ciências da computação (CARUANA; FREITAG, 1994; LANGLEY, 1994; GUYON; ELISSEEFF, 2003).

Em processos industriais, a seleção de variáveis, segundo Anzanello (2009), pode ser justificada por três aspectos: i) um modelo composto por elevado número de variáveis pode apresentar aderência satisfatória sobre os dados históricos, porém não oferece garantias em termos de predição (devido ao *overfitting*) e classificação (devido ao ruído de variáveis menos relevantes); ii) a identificação de variáveis de forma empírica é frequentemente sujeita a erros; e iii) modelos reduzidos são preferencialmente utilizados por demandarem menor tempo de análise e possuírem menor complexidade.

Em cenários industriais, existem dois objetivos principais para as sistemáticas de seleção de variáveis: *(i)* predição, em que o objetivo é encontrar um conjunto de variáveis independentes que viabilizam melhor predição da variável dependente, como em Wold *et al.* (2001), Gauchi e Chagnon (2001) e Pereira *et al.* (2011); e *(ii)* classificação, em que o objetivo é encontrar o conjunto de variáveis independentes que melhor discriminam novas observações entre classes, como em Urtubia *et al.* (2007), Anzanello (2009) e Anzanello *et al.* (2012). As proposições deste artigo estão alinhadas com a segunda frente, uma vez que a

correta classificação de bateladas, especialmente nos primeiros estágios de produção, permite o ajuste dos parâmetros para correção de inconsistências, podendo inclusive indicar o remanejamento da batelada para outro destino (ANZANELLO, 2009).

Este artigo propõe um novo método de seleção de variáveis para classificação de bateladas produtivas em duas categorias. O método consiste em quatro passos: (1) separar os dados originais em porções de treino e teste; (2) aplicar a ACP no banco de treino para a extração de informações a respeito da influência das variáveis na variabilidade do processo; através dos parâmetros da ACP é gerado um índice de importância de variáveis; (3) utilizar a ferramenta Máquina de Suporte Vetorial no banco de treino para calibrar um hiperplano separador de classes e classificar as amostras do banco de treino; remover a variável com menor índice de importância e repetir este procedimento iterativo até restar apenas uma variável; e (4) verificar a distância dos resultados dos subconjuntos em relação a um hipotético ponto ótimo. O subconjunto cujo resultado estiver mais próximo deste ponto é considerado como aquele que produz o hiperplano separador com melhor resultado. Por fim, classificar as amostras do banco de teste e verificar a acurácia do método. O método proposto é comparado à técnica OUVV, encontrada em Caruana e Freitag (1994), em termos da acurácia, quantidade de variáveis removidas e tempo de processamento dos métodos.

O restante do artigo é organizado como segue: a seção 2 apresenta um mapeamento de métodos de seleção de variáveis e os fundamentos das ferramentas utilizadas no método proposto, ACP e MSV. Na seção 3 é mostrado o método proposto. Na quarta seção são apresentados os bancos de dados utilizados e os resultados da aplicação da metodologia nestes bancos. Na quinta seção são trazidas as conclusões oriundas deste estudo e propostas para trabalhos futuros.

3.2 Fundamentação Teórica

3.2.1 Princípios teóricos da seleção de variáveis

Liu e Yu (2005) afirmam que existem quatro passos básicos para um método de seleção de variáveis: geração de subconjuntos, avaliação dos subconjuntos, critério de parada e validação do resultado. O fluxograma dos passos é mostrado na Figura 3.1.

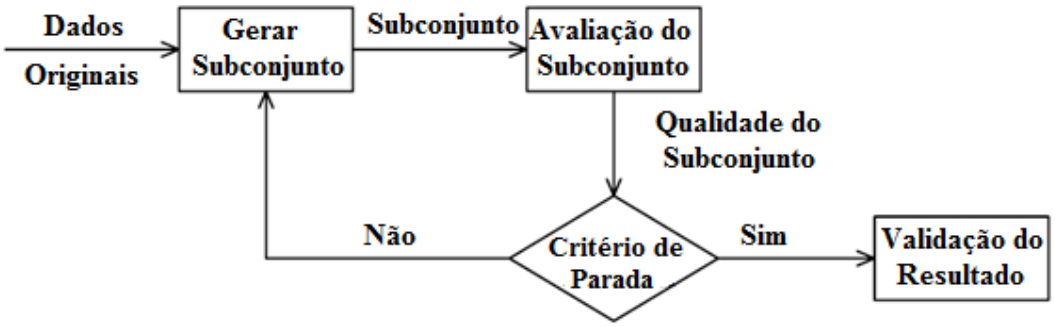

Figura 3.1 - Fluxograma para métodos de seleção de variáveis
Fonte: Liu e Yu (2005)

A geração de subconjuntos é um processo de busca apoiado em uma heurística, em que cada passo cria um subconjunto. Esta heurística deve inicialmente definir um ponto de início e direção da busca, que pode iniciar com um conjunto vazio e acrescentar variáveis (método *forward*), iniciar com o conjunto original de variáveis e removê-las (método *backward*), ou iniciar em ambos os métodos e remover e adicionar variáveis simultaneamente (método bidirecional ou *stepwise*). Além disto, é necessário definir o método de busca, que pode ser completo (verificando todas as permutações de subconjuntos de variáveis), sequencial (onde as variáveis são adicionadas ou removidas sistematicamente, em grupos ou individualmente) ou aleatório (onde as variáveis são adicionadas ou removidas aleatoriamente, em grupos ou individualmente).

Para a avaliação dos subconjuntos, são utilizados dois critérios: independentes, tais como medidas de distância, informação, dependência e consistência, ou dependentes, como ferramentas classificadoras. Critérios dependentes tipicamente levam a melhores acurácias, porém necessitam de maior desempenho computacional, uma vez que é necessário aplicar a ferramenta em todos os subconjuntos (LIU; YU, 2005).

A geração de subconjuntos é realizada até atingir um pré-determinado critério de parada. Os critérios de parada mais utilizados são: completar a busca, atingir algum limite (por exemplo, um número máximo de iterações), não haver melhora no processo com o passo subsequente, ou o alcance de um subconjunto com resultados suficientemente bons (DASH; LIU, 1997; BURGES, 1998; LIU; YU, 2005).

A validação do método pode ser verificada através de conhecimento posterior dos dados selecionados. Porém, como não há uma expectativa de quais dados serão selecionados, tipicamente tal conhecimento não existe. Portanto uma alternativa é a comparação da acurácia gerada pelas variáveis selecionadas frente à acurácia obtida quando todas as variáveis são

utilizadas. Para maior aprofundamento dos conceitos, recomenda-se a leitura de Dash e Liu, (1997), Guyon e Elisseeff (2003) e Li e Yu (2005).

3.2.2 Análise de Componentes Principais (ACP)

A ACP, inicialmente formulada por Pearson (1901) e aprimorada por Hotelling (1933), é uma ferramenta quantitativa multivariada que permite a redução de dimensionalidade e extração de características de dados com elevados níveis de ruído, correlacionados e dados espúrios. Em tese, qualquer conjunto de dados pode ser simplificado pela ACP (WOLD, 1987; YANG; WANG, 1999; ESBENSEN, 2002; DHARMARAJ *et al*., 2006; ABDI, 2010).

A ACP tem como objetivo substituir um conjunto de variáveis correlacionadas, $x_1, x_2, \dots, x_N$, por um conjunto de variáveis ortogonais, $t_1, t_2, \dots, t_N$, chamadas de componentes principais, que mantêm a variabilidade original dos dados (WOLD, 1987; NETO; MOITA, 1998; ABDI, 2010). Estas novas variáveis são combinações lineares das variáveis originais e resultam da equação (1),

$$t_i = \sum_{j=1}^{N} w_{ij}\, x_j, \quad \text{onde } i=1,2,\dots,N \qquad (1)$$

onde w_{ij} são os pesos de cada variável x_j em determinada combinação linear, componente, t_i. Segundo Neto e Moita (1998), cada componente t_i possui uma variância λ_i associada, de tal forma que os componentes são arranjados em ordem decrescente de variância.

Apesar de esta técnica gerar *N* componentes, não há um consenso de quantos componentes principais devem ser retidos, tornando esta escolha subjetiva. Adler e Yazhemsky (2010) recomendam a retenção de componentes que correspondam a, no mínimo, 80% da variabilidade. Já Cullen e Crouch (1997) e Tenenbaum *et al.* (2000) recomendam a análise visual das variâncias dos componentes principais em um gráfico, retendo os componentes de acordo com o "ponto de cotovelo" (a partir do qual não se verifica incrementos significativos de explicação da variância com adição de componentes extras).

3.2.3 Máquina de Suporte Vetorial (MSV)

Criada por Vapnik (1995), a MSV é uma técnica de aprendizagem computacional (sistemas que reconhecem padrões em dados) originalmente utilizada em problemas de

classificação em dois grupos (CORTES; VAPNIK, 1995). Para realizar tal classificação, a técnica utiliza um hiperplano para definir o limite de separação entre estes grupos. A calibração do hiperplano é realizada utilizando dois subplanos auxiliares, um em cada lado do hiperplano, de forma que a distância dos dois subplanos ao hiperplano seja máxima, penalizando os casos onde há amostras separadas incorretamente. O objetivo de tal hiperplano é mapear a origem de um plano *n*-dimensional, ou até com infinitas dimensões, onde a categorização dos dados seja mais simples (ANZANELLO, 2009; LUTS *et al*., 2010; CHEN *et al*., 2011). Maiores informações sobre a ferramenta são encontradas em Vapnik (1995), Burges (1998), Cristianini e Shawe-Taylor (2000), Hastie *et al*. (2005) e Huang e Wang (2006).

Ao contrário de outros métodos classificadores que se utilizam de limites de separação, a classificação da MSV é determinada por um pequeno número de *H* amostras do banco de treino, chamadas de Suportes Vetoriais (SV), que se localizam próximas ao limite de separação (ZOMER *et al*., 2004). Definindo os grupos de análise como 1 e -1 para classificar uma amostra x_i do banco de teste, Akay (2009) e Dixon e Brereton (2009) determinam que deve ser examinado o resultado da equação (2)

$$y_i = \text{sinal}\left(\sum_{h=1}^{H} \alpha_h * y_h * s_h * x_i + b\right) \qquad (2)$$

onde α_h é o Multiplicador de Lagrange, y_h é a classe (± 1) do SV correspondente, s_h é a amostra de cada *h* SV e *b* é o viés do parâmetro.

Transformações podem ser executadas nos dados originais, visando aumentar o poder de separação da MSV. Estas transformações movem os pontos no espaço original, facilitando a criação do limite de separação. Estas funções são chamadas de kernel, sendo a transformação RBF (ou Gaussiana) a mais utilizada (ANZANELLO, 2009; CHANG, 2010).

3.3 Metodologia para seleção de variáveis

Muitos modelos de seleção de variáveis incluem algum ranqueamento das variáveis como principal diretriz devido a sua simplicidade e bons resultados empíricos (GUYON; ELISSEEFF, 2003). Quando ranqueadas, a seleção de variáveis torna-se um procedimento de busca em etapas, em que a cada etapa é gerado um candidato a subconjunto recomendado das variáveis originais. Cada novo subconjunto é comparado com o melhor subconjunto anterior por um critério de avaliação, substituindo-o como eleito quando considerado melhor. Este

procedimento é repetido até que seja satisfeito algum critério de parada (LANGLEY, 1994; LIU; YU, 2005). Estruturado a partir destes conceitos, o método de Remoção Ordenada por Importância (ROI) tem seus passos apresentados nesta seção.

Passo 1: Separar os dados históricos em bancos de treino e teste

Tomando como base a matriz *X*, que contém todas as observações do banco de dados, criam-se randomicamente duas matrizes *Tr* (treino) e *Ts* (teste), onde *Tr* + *Ts* = *X*. A matriz *Tr* é utilizada para encontrar o subconjunto de variáveis que melhor classifica o processo, enquanto *Ts* é utilizada para a verificação de acurácia do método (representando novas observações). A quantidade de amostras em cada matriz de treino e de teste é definida pela proporção 3:2, recomendada por Chong *et al.* (2007).

Passo 2: Gerar o IIV

O IIV é utilizado para ordenar a remoção de variáveis não importantes ou que prejudiquem a classificação, substituindo a necessidade de enumerar todas as possibilidades existentes (o que é inviável em bancos de dados com muitas variáveis). O principal objetivo de eliminar as variáveis através de tal sistemática é prevenir correlações espúrias que podem ser tratadas como informações relevantes (WESTAD *et al*., 2003).

Segundo Neto e Moita (1998), a variável com maior peso nos primeiros componentes são as mais importantes na variabilidade do processo. Baseado nisso, o IIV proposto neste artigo parte dos pesos w_{ij} das *N* variáveis nos *p* componentes principais retidos, os quais são ponderados pela variância λ_i associada a cada componente. Como a análise é baseada na influência, tanto positiva quanto negativa, de w_{ij} nos componentes retidos, estes valores são somados na forma absoluta. O IIV associado a cada variável *j* é gerado pela equação (3).

$$IIV_j = \sum_{i=1}^{p} |w|_{ij} * \lambda_i \quad (3)$$

Assume-se que uma variável com menor IIV explique menos variabilidade do processo, ou é redundante a outra variável com maior IIV, o que pode resultar em uma análise menos precisa. Anzanello *et al*. (2011), apresentam um IIV semelhante, onde o índice é constituído pela soma absoluta dos pesos. Verifica-se, no entanto, que dentre os componentes principais retidos, pode haver variáveis com pesos desproporcionais à sua variabilidade

associada. A ponderação por λ_i diminui esta influência nos casos em que são retidos mais de um componente principal.

Passo 3: Classificar as observações do banco de treino em duas classes e remover variáveis irrelevantes

Nesta etapa as bateladas são classificadas como "conformes" ou "não conformes" por meio da Máquina de Suporte Vetorial (MSV). Inicialmente, as observações pertencentes à matriz *Tr* são utilizadas para calcular o limite de separação utilizado na classificação das observações.

Definido o limite de separação, classificam-se as amostras da porção de treino. Após classificadas todas as amostras, é calculada a acurácia de classificação do subconjunto com *n* variáveis (ACC_n) através da equação (4) e removida a variável com menor IIV da matriz *Tr*. Com essa nova matriz é calculado um novo limite de separação, com a dimensão apropriada, e novamente classificam-se as observações. As variáveis são removidas até restar apenas uma.

$$ACC_n = \frac{\text{Classificações corretas}}{\text{Total de amostras}} \qquad (4)$$

Passo 4: Determinar o melhor subconjunto de variáveis e classificar o banco de teste com as variáveis retidas

Para selecionar o subconjunto de variáveis que melhor classifica o processo, é utilizada uma simplificação da ferramenta existente em Anzanello *et al*. (2012). A ferramenta aqui utilizada consiste em encontrar o ponto, de coordenadas (acurácia,% variáveis retidas), com a menor distância em relação a um ponto hipotético ideal definido pelo usuário. Como as medidas de desempenho (acurácia e % de retenção de variáveis) estão no intervalo [0,1], é considerado como melhor cenário aquele onde apenas uma variável retida resulta em uma acurácia de 100%. O cálculo das distâncias é mostrado na equação (5)

$$d_n = \sqrt{(1 - ACC_n)^2 + (\frac{1-n}{N})^2} \qquad (5)$$

onde N é a quantidade original de variáveis, *n* é a quantidade de variáveis retidas, d_n é a distância do ponto com *n* variáveis retidas até o ponto ótimo e ACC_n é a acurácia de classificação com *n* variáveis retidas. O subconjunto de variáveis com menor distância ao

ponto hipotético ideal forma o conjunto considerado como melhor classificador do processo, como ilustrado na Figura 3.2.

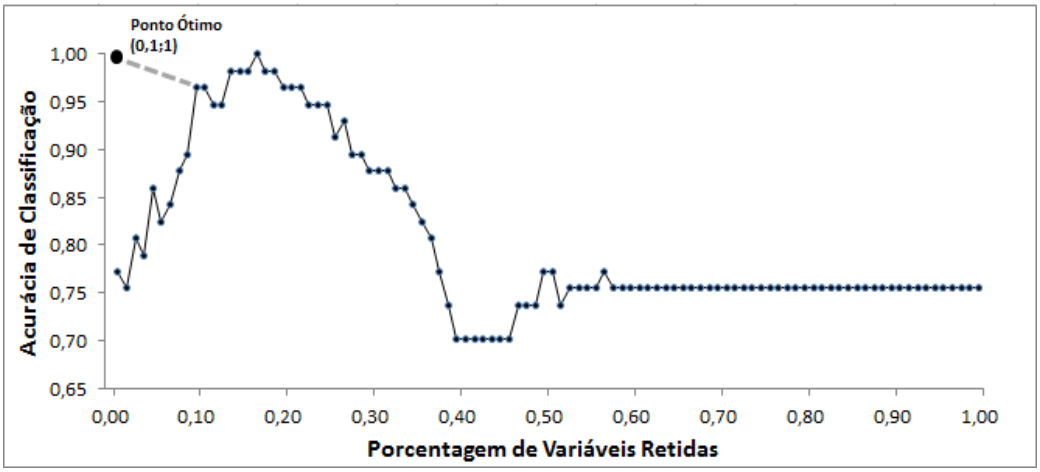

Figura 3.2 - Perfil hipotético de acurácia x porcentagem de variáveis retidas
Fonte: Os autores

Uma vez definido o melhor subconjunto de variáveis, classificam-se as observações do banco de teste, que representam novas bateladas não incluídas na geração do modelo. Verifica-se a acurácia do método por meio da equação (4).

Os resultados do método proposto são comparados com o método Omita Uma Variável por Vez (OUVV), técnica popular de seleção de variáveis. No OUVV, os dados originais são divididos na mesma proporção de treino e teste do método proposto. Para escolher qual a variável a ser removida, verifica-se a acurácia de classificação, por meio da MSV, do banco de treino com a remoção das variáveis uma a uma. A variável que, quando removida, apresentar maior acurácia será permanentemente excluída (visto que sua omissão do banco de dados eleva a acurácia de classificação, atestando sua limitada contribuição no processo). No caso de mais de uma variável, quando removida, levar à maior acurácia, é selecionada, aleatoriamente, uma entre elas. Repete-se este processo até restar apenas uma variável. O subconjunto de variáveis retidas é definido por meio da equação (5) e, utilizando estas variáveis, verifica-se a acurácia deste método.

3.4 Resultados

Os métodos são aplicados em quatro bancos de dados, compostos por *M* bateladas industriais e *N* variáveis de processo, aqui referidos como ADPN, GRANU, OXY e SPIRA. ADPN descreve a produção de um subcomponente na indústria de nylon; GRANU mostra os dados da produção de emulsões em um processo de fabricação de papel; OXY advém do

processo de obtenção de dióxido de titânio; SPIRA se refere à produção de antibióticos. Mais informações a respeito da origem dos bancos de dados são encontradas em Gauchi e Chagnon (2001).

A Tabela 3.1 mostra as dimensões dos bancos de dados e número de componentes principais retidos necessários para obter-se pelo menos 80% da variabilidade original após aplicação da ACP, com sua respectiva representação de variabilidade do processo.

Tabela 3.1 - Bancos de Dados e Componentes Principais retidos

Banco de dados	Número de variáveis	Número de observações (60-40)		Componentes Principais retidos	% de variabilidade retida
		Treino	Teste		
ADPN	100	43	28	3	83,99
GRANU	78	17	12	3	82,80
OXY	95	15	10	5	83,26
SPIRA	96	87	58	18	81,27

Com vistas à avaliação de desempenho do método, foram realizadas 300 repetições por meio da permuta aleatória de observações nos bancos de treino e teste, diminuindo assim a influência da presença de uma observação atípica no banco de teste. A análise foi realizada através do *software* Matlab® versão 7.8.0, em um computador com processador Intel® Core i7-2600, 3,40GHz e 4Gb de memória. Os códigos foram gerados pelos autores.

A Tabela 3.2 apresenta os resultados (acurácia média, desvio padrão da acurácia, média da porcentagem de variáveis retidas, desvio padrão da porcentagem de variáveis retidas e tempo de processamento) dos dois métodos, ROI e OUVV, e da MSV com todas as variáveis (sem seleção). A Tabela 3.2 mostra que a acurácia média dos bancos é igual a 53,2% quando todas as variáveis são consideradas. Utilizando os níveis definidos em Gomes (1985), os Coeficientes de Variação (CV) dos resultados de cada banco de dados são classificados como baixo (abaixo de 0,10), ou médio (entre 0,10 e 0,20), logo se pode afirmar que estas médias não contêm valores demasiadamente diferentes dos demais. Comparando os resultados da Tabela 3.2, pode-se afirmar que a utilização dos métodos de remoção de variáveis trouxe ganhos em comparação à classificação das amostras com todas suas variáveis. O ROI resultou em um ganho médio de 25,14% na acurácia, enquanto o OUVV aumentou a acurácia média em 16,43%. Também se observa que os CVs gerados pelo ROI não apenas estão classificados como baixo ou normal, como também são menores que os do OUVV (que também possui os CV baixos ou normais, à exceção do banco de dados GRANU que é classificado como alto). Esta comparação mostra que, além de serem mais confiáveis, os resultados do ROI não geram

resultados demasiadamente distantes da média. Tal robustez do método é desejada, uma vez que valores dispersos em relação à média podem prejudicar o controle do processo.

Tabela 3.2 - Resultados dos métodos ROI e OUVV e da MSV sem remoção de variáveis

Resultado	Banco de Dados	ROI	OUVV	Sem Remoção
Acurácia Média	ADPN	0,768	0,711	0,554
	GRANU	0,779	0,593	0,484
	OXY	0,859	0,801	0,610
	SPIRA	0,727	0,680	0,480
Desvio da Acurácia	ADPN	0,084	0,092	0,036
	GRANU	0,096	0,140	0,076
	OXY	0,128	0,138	0,076
	SPIRA	0,067	0,088	0,081
% de Variáveis Retidas	ADPN	10,34	6,54	-
	GRANU	4,62	5,56	-
	OXY	3,86	2,75	-
	SPIRA	11,05	9,25	-
Desvio da % de Variáveis Retidas	ADPN	4,15	1,74	-
	GRANU	3,05	1,89	-
	OXY	2,22	0,89	-
	SPIRA	2,84	2,50	-
Tempo (h)	ADPN	0,169	6,836	0,001
	GRANU	0,062	1,618	0,001
	OXY	0,067	3,566	0,001
	SPIRA	0,968	53,707	0,016

Durante a execução da OUVV, verificou-se que, durante as primeiras iterações, frequentemente ocorreu um empate entre todas as variáveis, tornando as primeiras eliminações aleatórias, porém não há como mensurar a influência deste fato nos valores finais. Apesar de obter uma acurácia menor, o OUVV necessita de menos variáveis para chegar a tal valor. Para avaliar qual método produz os melhores resultados, comparando acurácia e retenção de variáveis, é utilizada novamente a equação (5), definindo como melhor resultado aquele que estiver menos distante do ponto ótimo. Na Tabela 3.3 estão os valores destas distâncias.

Tabela 3.3 - Distância dos resultados até o ponto ótimo

	Distância	
	ROI	OUVV
ADPNN	0,250	0,294
GRANU	0,223	0,409
OXY	0,144	0,200
SPIRA	0,290	0,330

A Tabela 3.3 mostra que os resultados do ROI estão mais próximos do ponto ótimo em comparação aos do OUVV, portanto conclui-se que o ROI tem melhores resultados quando analisadas conjuntamente a acurácia e a quantidade de variáveis retidas.

Por fim, tem-se que o tempo de processamento demandado pelo ROI é significativamente menor que do OUVV em todos os bancos de dados (possuindo valores até 55 vezes menores). Já o tempo de processamento da MSV com todas as variáveis se mostrou muito baixo, uma vez que, ao contrário dos métodos de seleção de variáveis, as observações dos bancos de dados não são permutadas e a MSV é executada apenas uma vez.

3.5 Conclusão

Foi proposto neste artigo um novo método para selecionar variáveis que descrevem bateladas produtivas, Remoção Ordenada por Importância (ROI). Este método consiste em (1) separar os dados históricos em bancos de treino e teste; (2) aplicar a ACP nos dados e gerar o IIV por meio dos pesos relacionados às variáveis dos componentes retidos ponderados pela sua variância explicada; (3) classificar as amostras do banco de treino utilizando a MSV, calcular a acurácia, retirar a variável com menor IIV e repetir a classificação até restar apenas uma variável; e (4) definir o subconjunto de variáveis que melhor classifica o processo e verificar a acurácia do método utilizando as variáveis selecionadas para classificar o banco de testes.

Ao ser aplicado em quatro processos industriais, o método obteve um incremento médio de acurácia de 25,14%, quando comparado à classificação com todas as variáveis, com a retenção média de 7,47% das variáveis originais. Frente ao método OUVV, o ROI mostrou melhor desempenho e menor tempo de processamento.

Desdobramentos futuros incluem a criação de outras formas de ranqueamento das variáveis, utilizando técnicas que verifiquem a capacidade discriminante das variáveis, a dependência entre elas, ou ainda uma mistura destas duas vertentes. A extensão do método para o caso de múltiplas classes, onde as bateladas conformes podem ser inseridas em grupos de qualidades, também é sugerida.

3.6 Referências

ABDI, H; WILLIAMS, L.J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, v.2, n.4, p. 433-459, 2010.

ADLER, N.; YAZHEMSKY, E. Improving discrimination in data envelopment analysis: PCA–DEA or variable reduction. **European Journal of Operational Research**, v. 202 n. 1, p. 273-284, 2010.

AKAY, M.F. Support vector machines with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, n. 2, p. 3240-3247, 2009.

ANZANELLO, M. J. Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes. **Gestão & Produção**, v.16, n.4, p.526-533, 2009.

ANZANELLO, M.J.; FOGLIATTO, F.S.; ROSSINI, K. Data mining-based method for identifying discriminant attributes in sensory profiling. **Food Quality and Preferences**, v. 22, n. 1, p. 139-148, 2011.

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97-105, 2012.

BLUM, A.L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v.97, n.1, p.245-271, 1997.

BURGES, C.J.C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, v.2, n.2, p.121-167, 1998.

CARUANA, R.; FREITAG, D. Greedy attribute selection. **International Conference on Machine Learning**, p.28-36, 1994.

CHANG, Y.W.; HSIEH, C.J.; CHANG, K.W.; RINGGAARD, M.; LIN, C.J. Training and testing low-degree polynomial data mappings via linear SVM. **The Journal of Machine Learning Research**, v.99, p.1471-1490, 2010.

CHEN, H.L.; YANG, B.; LIU, J.; LIU, D.Y. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v.38, n.7, p.9014-9022, 2011.

CHONG, I.; ALBIN, S.L.; JUN, C. A data mining approach to process optimization without an explicit quality function. **IIE Transactions**, v.39, n.8, p.795-804, 2007.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v.20, n.3 p.273-297, 1995.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge university press, 2000.

CULLEN, T. F.; CROUCH, S. R. Multicomponent kinetic determinations using multivariate calibration techniques. **Microchimica Acta**, v.126, n.1-2, p.1-9, 1997.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent data analysis**, v.1, n.3 p.131-156, 1997.

DHARMARAJ, S.; HOSSAIN, M. A.; ZHARI, S.; HARN, G. L.; ISMAIL, Z. The use of principal component analysis and self-organizing map to monitor inhibition of calcium oxalate crystal growth by Orthosiphon stamineus extract. **Chemometrics and Intelligent Laboratory Systems**, v.81, n.1, p.21-28, 2006.

DIXON, S.; BRERETON, R.G. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. **Chemometrics and Intelligent Laboratory Systems**, v.95, n.1, p.1-17, 2009.

ESBENSEN, K.H. **Multivariate Data Analysis - In Practice**. 5ª ed. Oslo: CAMO Process AS, 2002.

EVANS, D.; CHAIX, B.; LOBBEDEZ, T.; VERGER, C.; FLAHAULT, A. Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology. **BMC Medical Research Methodology**, v.12, n.1, p.156, 2012.

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v.58, n.2, p.171-193, 2001.

GREENLAND, S. Modeling and variable selection in epidemiologic analysis. **American Journal of Public Health**, v.79, n.3, p.340-349, 1989.

GOMES, F.P. **Curso de estatística experimental**.11ª ed. Piracicaba: Livraria Nobel S.A., 1985.

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

HAPFELMEIER, A.; ULM, K. A new variable selection approach using Random Forest. **Computational Statistics and Data Analysis**, v.60, p.50-69, 2013.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.; FRANKLIN, J. The elements of statistical learning: data mining, inference and prediction. **The Mathematical Intelligencer**, v.27, n.2, p.83-85, 2005.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v.24, n.6, p. 417-441, 1933.

HUANG, C.L.; WANG, C.J. A GA-based feature selection and parameters optimization for support vector machines. **Expert Systems with Applications**, v.31, n.2, p. 231-240, 2006.

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

LANGLEY, P. Selection of relevant features in machine learning. **AAAI Fall Symposium on Relevance**, New Orleans, LA, p.140-144, 1994.

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

LUTS, J.; OJEDA, F.; Van de PLAS, R.; De MOOR, B.; Van HUFFEL, S.; SUYKENS, J.A. A tutorial on support vector machine-based methods for classification problems in chemometrics. **Analytica Chimica Acta**, v.665, n.2, p.129-145, 2010.

NETO, J.M.M.; MOITA, G.C. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, v.21, n.4, p. 467-469, 1998.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v.2, n.11, p.559-572, 1901.

PEREIRA, A.C.; REIS, M.S.; SARAIVA, P.M.; MARQUES, J.C. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. **Chemometrics and Intelligent Laboratory Systems**, v.105, n.1, p.43-55, 2011.

RUGGIERI, E.; LAWRENCE, C.E. On efficient calculations for Bayesian variable selection. **Computational Statistics and Data Analysis**, v.56, n.6, p.1319-1332, 2012.

TENENBAUM, J. B.; de SILVA, V.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v.290, n.5500, p.2319-2323, 2000.

URTUBIA, A.; PÉREZ-CORREA, J.R.; SOTO, A.; PSZCZOLKOWSKI, P. Using data mining techniques to predict industrial wine problem fermentation. **Food Control**, v.18, n.12, p.1512-1517, 2007.

VAPNIK, V. **The nature of statistical learning theory**. New York: Springer, 1995.

WESTAD, F.; HERSLETH, M.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v. 14, n. 5, p. 463-472, 2003.

WOLD, S.; SJOSTROM, M.; ERIKSEN, L. PLS-regression: a basic tool of chemometrics, **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109-130, 2001.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v.2, n.1, p.37-52, 1987.

YANG, T.; WANG, S. Robust algorithms for principal component analysis. **Pattern Recognition Letters**, v. 20, n. 9, p. 927-933, 1999.

ZOMER, S.; SÁNCHEZ, M.N.; BRERETON, R. G.; PAVON, J.L.P. Active learning support vector machines for optimal sample selection in classification. **Journal of Chemometrics**, v.18, n.6, p.294-305, 2004.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v.67, n.2, p.301-320, 2005.

4. Terceiro Artigo: Comparação de combinações de Índices de Importância de Variáveis e ferramentas classificatórias em um método *wrapper* de seleção de variáveis para classificação de bateladas produtivas

Alessandro Kahmann

Michel José Anzanello

Universidade Federal do Rio Grande do Sul

Resumo

Nas últimas décadas, houve uma significativa evolução das tecnologias para monitoramento de processos industriais, permitindo extrair e armazenar uma quantidade cada vez maior de dados. Tais dados são tipicamente impregnados por variáveis ruidosas ou altamente correlacionadas, reduzindo a eficiência de diversas ferramentas multivariadas de controle e monitoramento de processos. Para o estudo de bancos de dados com estas características, são utilizadas técnicas de mineração de dados, associadas a sistemáticas de seleção de variáveis, que buscam padrões de comportamento em dados históricos com o objetivo de prever comportamentos futuros. Neste artigo são comparadas diversas variações de um método do tipo *wrapper* de seleção de variáveis, com vistas à classificação de bateladas produtivas. O método consiste em separar os dados históricos em porções de treino e teste; na porção de treino, gera-se um Índice de Importância de Variáveis (IIV) que ordenará as variáveis de acordo com sua presumida capacidade discriminante; a cada iteração, classificam-se as observações da porção de treino e removem-se sistematicamente as variáveis; avaliam-se então os subconjuntos gerados através da distância Euclidiana dos resultados dos subconjuntos a um ponto hipotético ótimo. A classificação da porção de teste, utilizando apenas as variáveis retidas, determina a acurácia do método. Nas proposições deste artigo, são testados dois Índices de Importância de Variáveis (um baseado na dissimilaridade entre grupos e outro na Informação Mútua entre variáveis) e três ferramentas classificadoras (Máquina de Suporte Vetorial, K-Vizinhos Próximos e distância de Mahalanobis). Para verificar melhor as interações IIVs/ferramentas classificadoras, as mesmas variações do

método são aplicadas em bancos de dados reais e simulados com diferentes níveis de correlação e ruído.

Palavras-chave: Seleção de variáveis, Índices de Importância de Variáveis, Máquina de Suporte Vetorial, K-Vizinhos Próximos, Distância de Mahalanobis

Abstract

The recent technology evolution in industrial processes monitoring has allowed the storage of a large quantity of data. Such data are typically impregnated by noise and highly correlated variables, reducing the efficiency of several processes control tools. Data mining techniques have being used to study such databases and, in association with variable selection approaches, have enabled looking for behavior patterns in historical databases to forecast future behavior. This paper compares variations of a wrapper method of variable selection aimed at classifying production batches. The method initially splits the historical data into training and testing sets. Next, we generate a Variable Importance Index (IIV) that orders variables according to their presumed discriminate ability; at each iteration, samples of the training set are classified, and the less important variable is removed; the generated subsets are then evaluated by the Euclidian distance between the results and a hypothetical optimum. The testing set classification, using only the retained variables, determines the accuracy of the method. We test two VII (one based on the dissimilarity between groups, and another based on Mutual Information) and tree classification tools (Support Vector Machine, K-Nearest Neighbor and Mahalanobis distance). To define the best VII/classification tool, the combinations of indices and classifications tools are applied in simulated databases with different correlation and noise levels.

Keywords: Variable Selection, Variable Importance Index, Support Vector Machine, K-Nearest Neighbor, Mahalanobis Distance.

4.1 Introdução

A intensa competitividade nos mais diversos ramos de atuação incita as empresas a encontrarem soluções para se diferenciarem dos seus concorrentes e prosperarem no mercado.

Uma das formas de se obter diferenciação no mercado é por meio da comercialização de produtos com alta qualidade. Existem duas formas de abordagem para a qualidade do produto: (*i*) qualidade de projeto do produto, a qual está relacionada ao desenvolvimento de produtos que atendam às necessidades do cliente; e (*ii*) qualidade do processo de produção, a qual está relacionado à habilidade de se produzir bens que atendam às especificações técnicas previamente definidas (GARVIN, 2002; KLUG; MARSHALL, 2003; MARTINS, 2011).

O avanço de tecnologias para monitoramento de processos e armazenamento de dados faz com que controle de processo em ambientes industriais apoie-se em um número elevado de variáveis. Tal volume de dados pode comprometer a eficácia de ferramentas de análise, especialmente se impregnados por ruídos ou correlações elevadas (KOURTI; MACGREGOR, 1995; KETTANEH *et al.*, 2005; LIU; YU, 2005).

Dentro deste contexto, se torna necessário a utilização de técnicas de redução de dimensionalidade que permitam executar análises multivariadas sem a perda de informações importantes. Em processos industriais que operam com bateladas produtivas, o número de observações é tipicamente menor que o de variáveis, aumentando a dificuldade de análise (visto que diversas ferramentas falham nesta condição). Além disso, bancos de dados reduzidos são almejados para a viabilidade do monitoramento dos parâmetros do processo produtivo, permitindo identificar previamente mudanças de comportamento do processo, além de oferecer condições de classificar bateladas corretamente de acordo com as especificações desejadas (MARTIN *et al.,* 1999; GAUCHI; CHAGNON, 2001; ANZANELLO, 2009).

Mineração de dados é um processo computacional para reconhecimento de padrões em grandes bancos de dados, tendo como objetivo extrair informações relevantes destes bancos. Para tanto, são utilizadas técnicas de aprendizagem computacional, ferramentas estatísticas e sistemas de bancos de dados (GUYON; ELISSEEFF, 2003; HASTIE *et al*., 2005; KETTANEH *et al.*, 2005). A seleção de variáveis é uma das mais importantes, e frequentemente utilizadas, técnicas de mineração de dados. Ela tem por objetivo criar modelos de análise baseado apenas nas variáveis mais importantes do processo, removendo dados ruidosos, redundantes ou irrelevantes. Esta redução facilita a compreensão dos dados e aumenta a rapidez e acurácia de algoritmos de mineração de dados (BLUM; LANGLEY, 1997; GUYON; ELISSEEFF, 2003; LIU; YU, 2005; HAPFELMEIER; ULM, 2013; CHEN *et al.*, 2011; ULM, 2013). Tais fatos justificam o grande número de abordagens para seleção de variáveis em diversas áreas do conhecimento: Peña-Reyes e Sipper (1999), Akay (2009) e

Chen *et al.* (2011), para o diagnóstico de câncer de mama; Block *et al.* (1998), para prever deficiências cardiológicas; Anzanello *et al.* (2013), para identificar falsificações de remédios; Westad *et al.* (2003), para avaliar a opinião de consumidores; Ghose e Ipeirotis (2011), para mineração de texto e Rose Pehrsson *et al.* (2000), para criar um sistema de detecção de incêndio.

Segundo Saeys *et al.* (2007), a seleção de variáveis, pode ser justificada por três aspectos: i) evitar o *overfitting* e melhorar o desempenho do modelo; ii) fornecer modelos mais rápidos e com melhor relação custo-eficiência; iii) ganhar uma compreensão mais profunda a respeito do processo que gerou o dado. Em cenários industriais, sistemáticas de seleção de variáveis se dividem em dois objetivos principais: i) predição de uma ou mais variáveis dependentes, como em Wold *et al.* (2001), Gauchi e Chagnon (2001) e Pereira *et al.* (2011); e ii) classificação de novas observações entre classes, como em Urtubia *et al.* (2007), Chen *et al*. (2011) e Tian *et al*. (2013).

Gauchi e Chagnon (2001) comparam métodos de seleção de variáveis para predição através de modelos *filter* (quando a avaliação das variáveis a serem selecionadas ocorre através de uma medida, como por exemplo, distância entre as classes). Este artigo, por sua vez, comparará combinações de ferramentas multivariadas em um método de seleção de variáveis, de modelo *wrapper* (quando a avaliação das variáveis a serem selecionadas ocorre através de um modelo preditivo), com propósito de classificação de bateladas produtivas em duas categorias. O método apoia-se em quatro passos: (1) separar os dados originais em porções de treino e teste; (2) gerar um Índice de Importância de Variáveis com as observações da porção de treino; (3) utilizar uma ferramenta classificadora para classificar as observações do banco de treino; remover a variável com menor índice de importância e repetir este procedimento até restar apenas uma variável; (4) definir o melhor subconjunto de variáveis classificatórias verificando qual está mais próximo de um ponto ótimo hipotético. Por fim, classificar as observações do banco de teste e verificar a acurácia do método. Mais informações sobre os tipos de métodos de seleção de variáveis são encontrados em Guyon e Elisseeff (2003), Liu e Yu (2005) e Saeys *et al.* (2007).

Dentro deste método são testadas seis combinações entre dois IIVs (um baseado na dissimilaridade entre os grupos e outro na Informação Mútua entre as variáveis) e três ferramentas classificadoras (Máquina de Suporte Vetorial, K-Vizinhos Próximos e distância de Mahalanobis). Estas combinações são testadas em dados reais e simulados.

O restante do artigo está estruturado como segue: na seção seguinte são apresentados os Índices de Importância de Variáveis utilizados. Na terceira seção são apresentadas as ferramentas classificadoras. Na quarta seção é detalhado o método, explicando sua operacionalização. Na quinta seção é detalhada a simulação de dados para comparação dos métodos. Na sexta seção são apresentados os resultados das aplicações das variações do método em dados reais e simulados. Na sétima seção são mostradas as conclusões oriundas deste estudo.

4.2 Índices de Importância de Variáveis

Segundo Guyon e Elisseeff (2003), muitos métodos de seleção de variáveis incluem algum tipo de ordenação das variáveis, com o intuito de auxiliar a busca das variáveis que melhor descrevem o processo, por sua simplicidade e bons resultados empíricos. Diversos trabalhos, entre eles Rakotomamonjy (2003), Westad *et al.* (2003), Luts *et al.* (2010), Tian *et al.* (2013) e Hapfelmeier e Ulm (2013), utilizam algum método de ordenamento das variáveis para auxiliar a seleção.

Neste artigo o ordenamento das variáveis se dá através de um Índice de Importância de Variáveis (IIV), substituindo a necessidade de enumerar todas as possibilidades existentes (o que é inviável em bancos de dados com muitas variáveis).

4.2.1 Índice de Importância de Variáveis por Dissimilaridade

A distância de Bhattacharyya mede a similaridade de duas distribuições de probabilidade, tanto discretas quanto contínuas (BHATTACHARYYA, 1943). Sua atribuição original é encontrar a similaridade entre dois conjuntos de observações de dimensão *n*, mas em Coleman e Andrews (1979), a distância de Bhattacharyya entre os grupos *r* e *s* é simplificada para a soma da similaridade de uma variável por vez. O cálculo da similaridade de uma variável *j* por vez é mostrada na equação (1)

$$B_j(r,s) = \frac{1}{4} + \ln\left(\frac{1}{4}\left(\frac{\sigma_{rj}^2}{\sigma_{sj}^2} + \frac{\sigma_{sj}^2}{\sigma_{rj}^2} + 2\right)\right) + \frac{1}{4}\left(\frac{\left(\mu_{rj} - \mu_{sj}\right)^2}{\sigma_{rj}^2 + \sigma_{sj}^2}\right) \quad (1)$$

onde σ^2e μ são a variabilidade e a média da *j*-ésima variável dos grupos *r* e *s*, respectivamente.

A Figura 4.1 auxilia na compreensão do comportamento da distância de Bhattacharyya para uma variável por vez. Quando verificam-se variâncias iguais, mas médias diferentes, como na Figura 4.1(a), o primeiro termo é igual à zero e o segundo termo diferente de zero. Na Figura 4.1(b), percebe-se o caso contrário, onde as médias são iguais, porém as variabilidades são diferentes. Neste caso, o segundo termo será igual à zero, porém o primeiro será positivo. Este caso só apresentará uma distância alta se as variabilidades forem significativamente diferentes. Já no terceiro caso, mostrado na Figura 4.1(c), tem-se que tanto as médias quanto as variabilidades são diferentes, fazendo com que nenhum dos termos seja zero. Adicionalmente, pode-se afirmar que casos com variabilidade pequena e médias significativamente diferentes representam os casos onde há maior dissimilaridade entre as variáveis (COLEMAN; ANDREWS, 1979).

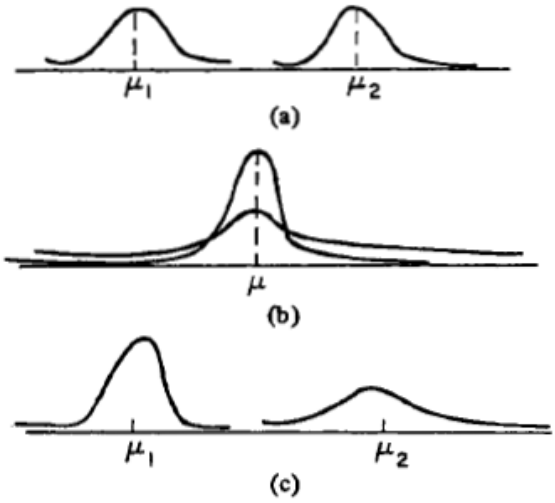

Figura 4.1 - Distância de Bhattacharyya para uma variável por vez. (a) Variâncias iguais e médias diferentes. (b) Médias iguais e variâncias diferentes. (c) Médias e Variâncias diferentes
Fonte: Coleman e Andrews (1979)

Segundo Liu e Yu (2005), na seleção de variáveis deseja-se encontrar o nível de similaridade de uma variável com sua classe. Considerando-se o caso de classificação em dois grupos, percebe-se que, ao buscar a variável de maior similaridade com sua classe, pode-se alternativamente buscar a variável com maior dissimilaridade com a outra classe. Sustentado nestes conceitos, o Índice de Importância de Variáveis por Dissimilaridade (IIV-D) calcula a distância de Bhattacharyya entre os dois grupos para cada variável individualmente. Segundo Coleman e Andrews (1979), uma variável com maior distância de Bhattacharyya é considerada como uma melhor separadora de classes, portanto deve ser mantida em detrimento de variáveis com menor distância.

4.2.2 Índice de Importância de Variáveis baseado na Informação Mútua

A Informação Mútua (IM) entre duas variáveis aleatórias é uma medida de dependência entre estas, ou seja, uma quantificação adimensional da informação que uma variável aleatória possui a respeito de outra. Esta quantificação pode ser pensada também como a redução da incerteza de uma variável aleatória por meio do conhecimento prévio de outra variável aleatória (POMPE, 1993; RACHOW *et al.,* 2011; LONG *et al.,* 2012). Segundo Pompe (1993) e Rodriguez-Rosario *et al.* (2008), a IM entre duas variáveis aleatórias é sempre não negativa, onde um valor alto ou pequeno indica que as variáveis são muito ou pouco relacionadas, respectivamente, ou independentes no caso em que a IM é nula.

O fundamento matemático da Informação Mútua é agora apresentado: Shannon (1948) traz o conceito de entropia, medida de incerteza de um evento X, ou H(X), e de entropia condicional, medida de incerteza de um evento X dado a observação prévia de um evento Y, ou H(X|Y). Baseado nestes conceitos, Long *et al.* (2012) afirmam que a IM entre dois eventos X e Y é descrita pela equação (2).

$$I(X;Y) = H(X,Y) - H(Y|X) - H(X|Y) \qquad (2)$$

As relações H(X), H(Y), H(X,Y), H(X|Y), H(Y|X) e I(X;Y) podem ser expressas em um diagrama de Venn, como mostrado na Figura 4.2. Note que I(X;Y) corresponde à intersecção da informação em X e da informação em Y.

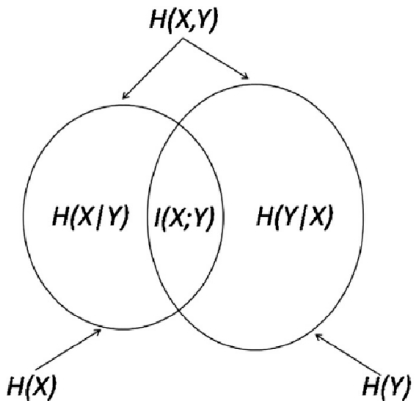

Figura 4.2 - Diagrama de Venn para relação entre entropia e IM
Fonte: Long et al. (2012)

O Índice de Importância de Variável baseado na Informação Mútua (IIV-IM), parte do pressuposto que um par com alta informação mútua implica que uma variável explica muito da variabilidade da outra variável. Como a IM é sempre não negativa, tomando um conjunto

de *N* variáveis de uma matriz *A* e calculando a IM de uma variável *j* com todas as variáveis restantes do conjunto, gera-se uma quantificação da informação relativa da variável *j*. Generalizando, o IIV-IM associado a cada variável *j* de um conjunto de *N* variáveis é dado pela equação (3).

$$IIV\text{-}IM_j = \sum_{l=1}^{N} \mathrm{I}(A_j; A_l) : j \neq l \qquad (3)$$

Desta forma, conclui-se que uma variável com alto IIV-IM possui grande quantidade de informações a respeito de outras variáveis, portanto é mais importante para a compreensão do processo quando comparada a uma variável com baixo IIV-IM.

4.3 Ferramentas de classificação: Máquina de Suporte Vetorial, K-Vizinhos Próximos e Distância de Mahalanobis

A primeira ferramenta de classificação é a Máquina de Suporte Vetorial (MSV). A MSV é uma técnica de aprendizagem computacional utilizada em problemas de classificação em dois grupos (CORTES; VAPNIK, 1995). Conforme ilustrado na Figura 4.3, a MSV cria um hiperplano separador entre as observações das classes, aqui chamadas de -1 e 1. Este plano é calibrado utilizando dois subplanos auxiliares, um em cada lado do hiperplano, de forma que a distância do hiperplano para os subplanos seja máxima, penalizando os casos onde elementos são deslocados para a classe incorreta. O objetivo de criar tal hiperplano é mapear a origem de um plano *n*-dimensional onde a categorização dos dados seja mais simples. Com vistas ao aumento do poder de categorização da MSV, transformações podem ser executadas nos dados originais previamente à calibração do hiperplano. Essas transformações, chamadas de kernel, movem os pontos no espaço original com o objetivo de facilitar a criação do limite de separação (RAKOTOMAMONJY, 2003; POLAT; GÜNES; 2007; ANZANELLO, 2009; CHANG, 2010; LUTS *et al*., 2010; CHEN *et al*., 2011). Maiores informações sobre a ferramenta e as transformações kernel são encontradas em Vapnik (1995), Burges (1998), Cristianini e Shawe-Taylor (2000), Hastie *et al*. (2005) e Huang e Wang (2006).

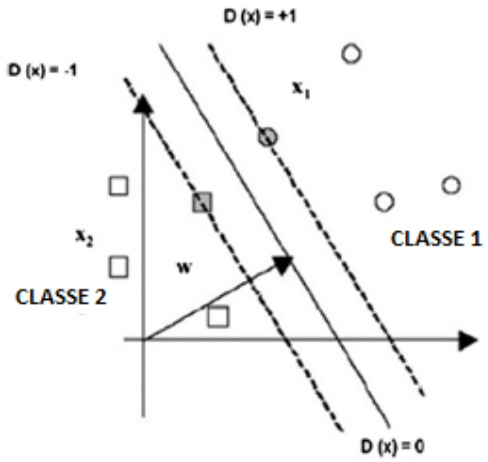

Figura 4.3 - A estrutura de uma MSV simples
Fonte: Polat e Günes (2007)

A segunda ferramenta de classificação, K-Vizinhos Próximos - KVP, classifica uma nova observação como pertencente à classe com maior incidência entre as *k* observações mais próximas. Considerando um conjunto multidimensional, com as observações deste conjunto divididas entre as classes -1 e 1, esta ferramenta classifica novas observações baseada em sua proximidade com as observações do conjunto. O algoritmo calcula a distância Euclidiana do ponto a ser classificado em relação aos pertencentes ao conjunto original. Tomando as classes das *k* observações mais próximas, a nova observação é classificada como pertencente à classe -1 se a maioria das observações pertencer a esta classe e classificada como 1 caso contrário (COVER; HART, 1967; CHAOVALITWONGSE, 2007; ANZANELLO, 2009). Esta ferramenta se destaca por sua simplicidade teórica, eficiência computacional e possuir apenas um parâmetro *k*, que pode ser definido através de validação cruzada.

A terceira ferramenta de classificação, a distância de Mahalanobis, primeiramente proposta em Mahalanobis (1939), é dada pela equação (4),

$$dm^2_{(amostra,grupo)} = (\mathbf{x} - \mu)^T \sum^{-1} (\mathbf{x} - \mu) \quad (4)$$

onde **x** é o vetor de variáveis da observação a ser classificada, μ é o vetor média dos componentes da matriz do grupo em comparação e ∑ é a matriz de covariância da matriz dos elementos do grupo em comparação. Apesar de basear-se na distância Euclidiana (que assume que cada variável é igualmente importante e independente), a distância de Mahalanobis pondera as variáveis da amostra pela sua matriz de covariância. De tal forma, criam-se elipsoides ao redor do centro, determinando escalas de distâncias da observação analisada até o respectivo centro de cada grupo, como ilustrado na Figura 4.4. Para o caso de classificação,

uma nova observação é inserida na classe cuja distância de Mahalanobis for menor. (MAESSCHALCK *et al*., 2000; XIANG *et al*., 2008; DIXON; BRERETON, 2009)

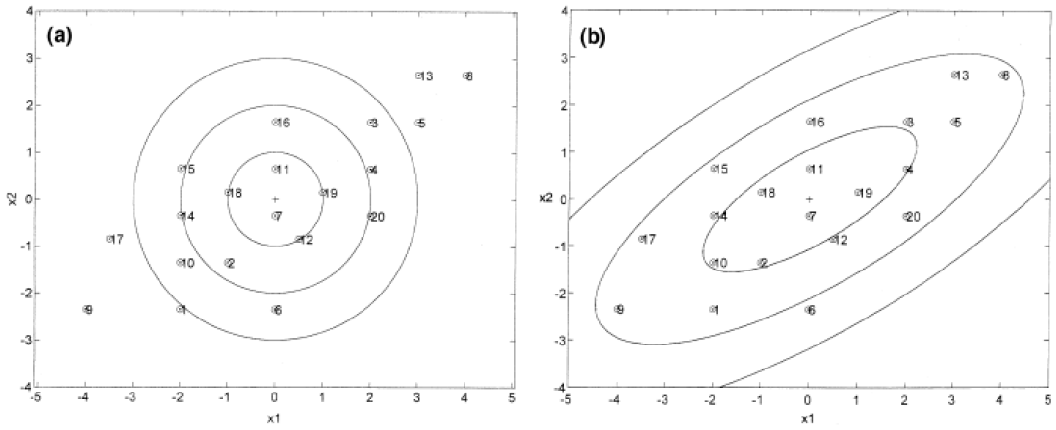

Figura 4.4 - (a) Distância de dados ao centro de acordo com a Distância Euclidiana; (b) Distância de dados ao centro de acordo com a Distância de Mahalanobis
Fonte: Maesschalck et al. (2000)

4.4 Metodologia para seleção de variáveis

Segundo Liu e Yu (2005), existem quatro passos básicos para um método de seleção de variáveis: geração de subconjuntos, avaliação dos subconjuntos, critério de parada e validação do resultado. Estruturado a partir deste conceito, o método proposto tem seus passos apresentados nesta seção, os quais são ilustrados no fluxograma da Figura 4.5.

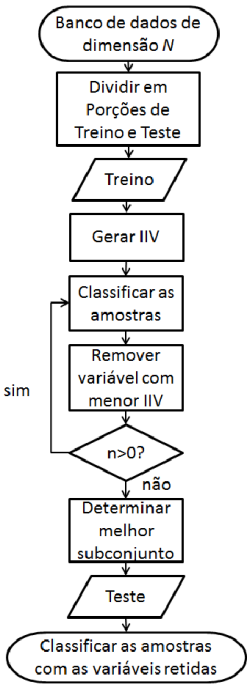

Figura 4.5 - Fluxograma da sistemática
Fonte: Os autores

Passo 1: Separar os dados históricos em bancos de treino e teste

Considere a matriz *A*, de tamanho *M* x *N*, que contém todas as observações do banco de dados. A partir de *A* são randomicamente criadas duas matrizes, *Tr* (treino) e *Ts* (teste), em que $Tr \cup Ts = A$. A matriz *Tr* é utilizada para encontrar o subconjunto de variáveis que melhor classifica as bateladas, enquanto *Ts* é utilizada para verificação de acurácia do método. A proporção de amostras em cada matriz de treino e de teste é 80:20, utilizada em Anzanello *et al.* (2012).

Passo 2: Gerar o IIV

Na porção de teste é gerado o IIV, que ordena as variáveis para remoção. Neste passo existem duas variações apresentadas na seção 2: IIV-D e IIV-IM, cada uma executada de forma independente.

Passo 3: Classificar as observações do banco de treino e remover variáveis irrelevantes

Nesta etapa, as bateladas são classificadas como "conforme" ou "não-conforme" por cada ferramenta classificadora apresentada na seção 3. As amostras pertencentes à matriz *Tr* são utilizadas para calibrar a ferramenta classificadora.

Após classificadas todas as amostras, é calculada a acurácia de classificação do subconjunto com *n* variáveis retidas (ACC_n) através da equação (5) e removida a variável com menor IIV da matriz *Tr*. Essa nova matriz é novamente classificada pela ferramenta classificadora. As variáveis são removidas até restar apenas uma.

$$ACC_n = \frac{\text{Classificações corretas}}{\text{Total de amostras}} \qquad (5)$$

Passo 4: Determinar o melhor subconjunto de variáveis e classificar o banco de teste com as variáveis retidas

Para selecionar o subconjunto de variáveis que melhor classifica o processo, é utilizada uma generalização da ferramenta existente em Anzanello *et al*. (2012). A ferramenta aqui utilizada consiste em encontrar o ponto, de coordenadas (% variáveis retidas, acurácia), com a menor distância em relação a um ponto hipotético ideal definido pelo usuário. Como as medidas de desempenho, acurácia e retenção de variáveis, estão no intervalo [0,1], é

considerado como melhor cenário aquele em que apenas uma variável retida resulta em uma acurácia de 100%. O cálculo das distâncias é mostrado na equação (6)

$$d_n = \sqrt{(1 - ACC_n)^2 + (\frac{1-n}{N})^2} \qquad (6)$$

onde n é a quantidade de variáveis retidas, d_n é a distância do ponto com n variáveis retidas até o ponto ótimo e ACC_n é a acurácia de classificação com n variáveis retidas. O subconjunto de variáveis com menor distância ao ponto hipotético ideal é considerado como o melhor classificador do processo, como ilustrado na Figura 4.6.

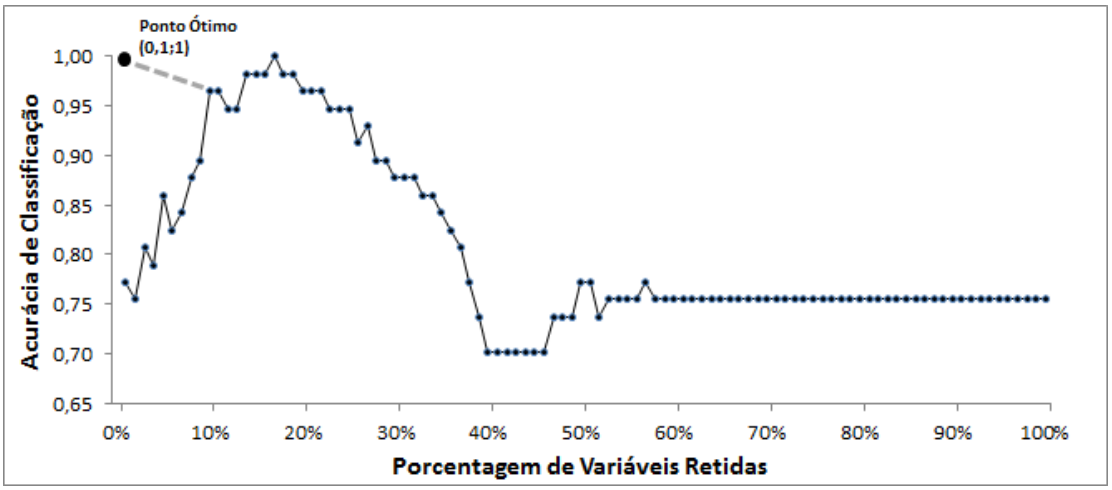

Figura 4.6 - Perfil hipotético acurácia x porcentagem de variáveis retidas
Fonte: Os autores

Uma vez definido o melhor subconjunto de variáveis, classificam-se as amostras do banco de teste, que representam novas bateladas não incluídas na geração do modelo. Verifica-se a acurácia desta classificação utilizando a equação (5).

4.5 Simulação de dados para comparação das variações do método

Os experimentos de simulação visam comparar o comportamento das variações do método em diferentes condições. Todas as simulações são baseadas nas características de um estágio de polimerização em um processo de produção de látex. Os dados originais possuem 117 variáveis de processo e 262 observações correspondentes a bateladas produtivas. Foram gerados dados baseados na equação de regressão dos dados originais, conforme equação (7)

$$y_i = \sum_{j=1}^{J} b_j x_{ij} + \varepsilon_i, j=1,2,...,J \quad (7)$$

onde $\varepsilon_i \sim N(0,\sigma^2)$. Os coeficientes b_j da regressão e a variância σ^2 do erro são obtidas dos dados reais. A estimação de σ^2 é realizada através da equação (8)

$$\sigma^2 = \frac{SQR}{n-1} \quad (8)$$

onde *SQR* é a soma quadrática residual dos dados reais e $m = 262$ observações. Para cada simulação i, os x_{ij} são gerados através de uma distribuição multinormal com média μ_j, para j=1,...,J variáveis do processo, e matriz de covariância Γ, ambos estimados do banco de dados original.

Para criar cenários diferentes, os fatores covariância e ruído são alterados. Os valores nominais de variância do erro e covariância são extraídos dos dados originais. Os diferentes níveis de covariância e variância do erro são mostrados na Tabela 4.1.

Tabela 4.1 - Níveis dos fatores da simulação

Fator	Níveis
Variância do erro	σ^2; $2{,}5\sigma^2$; $5\sigma^2$
Covariância das Variáveis	$0{,}5\Gamma$; Γ; 2Γ

4.6 Resultados

4.6.1 Dados reais

Os métodos propostos são aplicados em cinco bancos de dados, originalmente apresentados em Gauchi e Chagnon (2001). Tais bancos são compostos por *M* bateladas industriais descritas por *N* variáveis de processo, sendo referidos como ADPN, LATEX, OXY, SPIRA e GRANU. ADPN advém da produção industrial de adiponitrila, um subcomponente na indústria de nylon; LATEX descreve um estágio da polimerização da produção de látex, utilizado no revestimento de papel; GRANU mostra os dados da produção de emulsões em um processo de fabricação de papel; OXY foi obtido de um processo de obtenção de dióxido de titânio; SPIRA é obtido através do processo de fermentação usado na manufatura de um antibiótico. Mais informações a respeito dos bancos de dados são encontradas em Gauchi e Chagnon (2001).

A proporção da quantidade de dados nas porções de treino e teste será 80-20, respectivamente. A Tabela 4.2 mostra as dimensões dos bancos de dados.

Tabela 4.2 - Tamanho dos Bancos de Dados

Banco de dados	Número de variáveis no processo	Número de observações (80-20)	
		Treino	Teste
ADPN	100	57	14
LATEX	117	210	52
GRANU	78	23	6
OXY	95	20	5
SPIRA	96	115	29

Com vistas a uma melhor avaliação dos desempenhos, foram realizadas 300 repetiçõesem cada banco de dados. A cada repetição, as bateladas são aleatoriamente permutadas entre os bancos de treino e teste, mantendo-se a proporção definida. Objetiva-se não favorecer ou prejudicar o método com uma única separação dos bancos em treino e teste. Em cada variação do método (IIV e ferramenta classificadora), são utilizadas as mesmas 300 diferentes composições dos bancos. Os parâmetros das ferramentas são mostrados na Tabela 4.3; a análise foi realizada em Matlab® versão 7.8.0 e os códigos foram gerados pelos autores.

Tabela 4.3 - Parâmetros das ferramentas classificadoras

Banco de dados	*k* para a ferramenta KVP	Função Kernel (parâmetro)
ADPN	3	Polinomial (grau=3)
LATEX	3	RBF
GRANU	3	RBF
OXY	3	RBF
SPIRA	7	RBF

A Tabela 4.4 apresenta acurácia e desvio padrão das ferramentas classificadoras sem a remoção de variáveis, enquanto a Tabela 4.5 apresenta os resultados (acurácia média, desvio padrão da acurácia, média da porcentagem de variáveis retidas e desvio padrão da porcentagem de variáveis retidas) das seis variações do método proposto.

Tabela 4.4 - Resultados das classificações sem remoção

Resultado	Banco de dados	Ferramenta Classificadora			Resultado	Ferramenta Classificadora		
		MSV	KVP	Mahal		MSV	KVP	Mahal
Acurácia Média	ADPN	0,731	0,787	0,436	Desvio da Acurácia	0,104	0,097	0,108
	LATEX	0,707	0,825	0,797		0,055	0,052	0,054
	GRANU	0,346	0,798	0,590		0,121	0,145	0,170
	OXY	0,720	0,799	0,293		0,186	0,187	0,177
	SPIRA	0,656	0,740	0,406		0,076	0,071	0,067
	Média	0,632	0,790	0,505		0,108	0,110	0,115

Tabela 4.5 - Resultados das variações do método com dados reais

Resultado	Banco de Dados	IIV-D				IIV-IM			
		MSV	KVP	MAHAL	Média	MSV	KVP	MAHAL	Média
Acurácia Média	ADPN	0,817	0,759	0,810	0,795	0,764	0,702	0,773	0,746
	LATEX	0,816	0,789	0,806	0,804	0,762	0,720	0,796	0,759
	GRANU	0,712	0,743	0,701	0,719	0,753	0,789	0,727	0,756
	OXY	0,998	0,997	0,995	0,997	0,834	0,911	0,764	0,836
	SPIRA	0,765	0,730	0,746	0,753	0,703	0,642	0,691	0,679
	Média	0,822	0,804	0,812	0,813	0,763	0,753	0,750	0,755
Desvio da Acurácia	ADPN	0,097	0,102	0,097	0,099	0,110	0,112	0,109	0,110
	LATEX	0,053	0,070	0,058	0,060	0,051	0,063	0,058	0,057
	GRANU	0,164	0,152	0,168	0,161	0,148	0,153	0,180	0,160
	OXY	0,035	0,023	0,043	0,034	0,179	0,134	0,178	0,163
	SPIRA	0,070	0,091	0,072	0,078	0,084	0,088	0,082	0,084
	Média	0,084	0,088	0,088	0,086	0,114	0,110	0,121	0,115
% de Variáveis Retidas	ADPN	6,20	2,96	5,09	4,75	9,49	4,88	8,54	7,64
	LATEX	6,89	4,72	10,45	7,35	9,03	7,29	12,04	9,45
	GRANU	8,86	4,67	10,26	7,93	6,55	2,57	8,23	5,78
	OXY	1,05	1,06	1,07	1,06	2,09	1,06	4,86	2,67
	SPIRA	5,87	4,62	8,66	6,38	7,35	5,08	11,41	7,95
	Média	5,77	3,61	7,10	5,49	6,90	4,18	9,02	6,70
Desvio da % de Variáveis Retidas	ADPN	1,87	2,14	2,35	2,12	0,91	3,72	2,06	2,23
	LATEX	0,99	1,81	2,23	1,68	0,74	2,08	1,29	1,37
	GRANU	4,48	2,21	4,60	3,76	3,32	2,43	2,32	2,69
	OXY	0,00	0,06	0,20	0,09	1,05	0,10	1,41	0,85
	SPIRA	0,68	2,43	2,35	1,82	0,52	1,49	2,61	1,54
	Média	1,60	1,73	2,35	1,89	1,31	1,96	1,94	1,74

Em relação aos índices de importância das variáveis, o IIV-D conduziu a melhores resultados quando comparado ao IIV-IM, em termos de acurácia média e porcentagem média de variáveis retidas, com exceção do banco de dados GRANU. Enfatiza-se a elevada acurácia no banco OXY, onde o IIV-D identificou uma única variável capaz de definir com 100% de acurácia a categoria das amostras independentemente da ferramenta classificatória (os casos que não atingiram 100% de acurácia foram os que o método conduziu à retenção de mais de uma variável). Comparada às classificações sem remoção de variáveis, a remoção ordenada pelo IIV-D aumentou a acurácia média em 17% frente à acurácia obtida pela utilização de todas as variáveis, retendo em média 5,49% das variáveis originais, enquanto a remoção ordenada pelo IIV-IM aumentou a acurácia média em 11%, retendo em média 6,7% das variáveis originais. Um fato que chama a atenção é a KVP ter uma acurácia maior sem a remoção de variáveis, à exceção de quando aplicada no banco de dados OXY.

Sob as mesmas condições, as ferramentas classificatórias chegaram a resultados distintos. Considerando a média nos IIVs, a KVP conduziu à melhor acurácia nos bancos de dados com menos observações (OXY, 25 observações, e GRANU, 29 observações), enquanto a MSV conduziu à melhor acurácia nos bancos de dados com mais observações (ADPN, 71 observações, LATEX, 262 observações, e SPIRA, 144 observações). Para a retenção de variáveis, a KVP apresentou melhores resultados sob todas as condições.

Neste ponto, ressalta-se ainda uma característica negativa da distância de Mahalanobis. Bohling *et al*. (1998) citam que matrizes de covariância podem ser não-invertíveis, tornando necessária a utilização de uma matriz pseudo-inversa no cálculo da distância. Porém, tal adaptação pode prejudicar a ferramenta. Outra limitação encontrada da distância de Mahalanobis é o *overfitting*, que ocorre quando o modelo inclui ruído ou erro para melhorar a classificação da porção de treino, perdendo assim sua capacidade de generalização. Segundo Saeys *et al.* (2007), o *overfitting* é uma das desvantagens de métodos do tipo *wrapper*. Foi identificado *overffiting* do método apenas quando utilizada a distância de Mahalanobis, porém isso não significa que o mesmo não aconteça nas outras variações.

4.6.2 Dados Simulados

As seis variações propostas são executadas nos bancos de dados simulados propostos na Seção 4.5. Com base na Tabela 4.6, percebe-se perda de qualidade do método quando a correlação e o ruído aumentam. Como esperado, a acurácia diminui e a quantidade de variáveis retidas aumenta, assim como a variabilidade dos resultados cresce. Porém, um resultado que chama a atenção é a robustez da distância de Mahalanobis, independente do IIV utilizado. Quando a correlação aumenta, as ferramentas classificadoras KVP e MSV perdem seu poder de acurácia (apesar de não necessitar reter mais variáveis para isso), ao contrário da distância de Mahalanobis; em contrapartida, essa última necessita reter mais variáveis para isso. O IIV-IM produziu bons resultados em dados pouco correlacionados; o aumento da correlação, indiferentemente da ferramenta classificadora utilizada, reduz acurácia dos métodos. Peng *et al.* (2005) alertam que variáveis com alta Informação Mútua podem conter também alta redundância, justificando o fato de ferramentas que não lidam essencialmente com dados correlacionados, MSV e KVP, produzirem resultados menores conforme aumenta a correlação entre as variáveis.

Tabela 4.6 - Resultados das variações do método com dados simulados

Resultado	Correlação	Ruído	IIV-D			IIV-IM		
			MSV	KVP	MAHAL	MSV	KVP	MAHAL
Acurácia Média	0,5Γ	1σ²	0,844	0,849	0,840	0,849	0,790	0,850
		2,5σ²	0,845	0,849	0,848	0,852	0,830	0,843
		5σ²	0,814	0,814	0,803	0,817	0,762	0,810
	1Γ	1σ²	0,808	0,793	0,836	0,697	0,677	0,750
		2,5σ²	0,767	0,776	0,771	0,683	0,636	0,733
		5σ²	0,758	0,729	0,765	0,608	0,575	0,675
	2Γ	1σ²	0,741	0,727	0,781	0,637	0,575	0,745
		2,5σ²	0,782	0,770	0,789	0,647	0,572	0,725
		5σ²	0,699	0,686	0,776	0,632	0,581	0,725
Desvio da Acurácia	0,5Γ	1σ²	0,048	0,049	0,052	0,048	0,051	0,052
		2,5σ²	0,051	0,052	0,054	0,052	0,060	0,052
		5σ²	0,054	0,062	0,056	0,051	0,053	0,052
	1Γ	1σ²	0,058	0,063	0,053	0,057	0,067	0,064
		2,5σ²	0,064	0,062	0,064	0,062	0,065	0,069
		5σ²	0,064	0,063	0,065	0,070	0,076	0,075
	2Γ	1σ²	0,061	0,070	0,061	0,065	0,075	0,068
		2,5σ²	0,070	0,057	0,060	0,072	0,081	0,074
		5σ²	0,071	0,071	0,062	0,068	0,079	0,081
% de Variáveis Retidas	0,5Γ	1σ²	5,58	2,31	7,68	8,54	3,60	8,93
		2,5σ²	5,60	2,48	7,17	9,51	3,44	8,86
		5σ²	6,48	2,68	9,01	9,81	3,11	9,95
	1Γ	1σ²	5,78	2,58	7,60	11,65	4,79	12,46
		2,5σ²	6,25	2,10	10,81	11,66	4,53	12,63
		5σ²	6,60	3,74	9,75	11,22	4,51	13,20
	2Γ	1σ²	5,91	3,63	10,75	10,37	4,81	12,65
		2,5σ²	5,58	2,76	10,51	10,26	5,71	12,40
		5σ²	6,16	5,27	10,85	10,28	7,92	13,38
Desvio da % de Variáveis Retidas	0,5Γ	1σ²	1,00	1,28	1,23	3,37	2,11	1,36
		2,5σ²	1,03	1,27	1,25	1,84	1,40	1,17
		5σ²	0,79	1,02	1,77	1,05	1,88	1,41
	1Γ	1σ²	0,57	1,96	1,54	1,01	2,01	1,28
		1σ²	0,57	1,71	1,65	0,99	2,64	1,50
		2,5σ²	0,86	2,58	1,80	0,70	3,43	1,74
		5σ²	0,54	1,60	1,55	0,65	1,78	1,72
	2Γ	1σ²	0,62	1,42	1,43	0,47	3,62	1,64
		2,5σ²	0,64	2,97	1,66	0,43	2,87	1,84
		5σ²	1,00	1,28	1,23	3,37	2,11	1,36

4.7 Conclusões

Neste artigo são utilizadas combinações de IIV, para ordenamento de remoção de variáveis, com ferramentas classificatórias em um método *wrapper* de seleção de variáveis

para classificação. Este método consiste em (1) separar os dados históricos em bancos de treino e teste; (2) gerar um IIV; (3) classificar as observações do banco de treino, calcular a acurácia e remover a variável com menor IIV, repetindo este processo até restar uma variável; (4) definir o subconjunto que melhor classifica o processo e verificar a acurácia do método utilizando as variáveis selecionadas para classificar as observações do banco de teste. Nas proposições deste artigo são utilizadas duas formas de IIVs; o IIV-D, baseado na distância de Bhattacharyya, e o IIV-IM, baseado na Informação Mútua das variáveis; e três ferramentas classificadoras com características distintas, MSV, KVP e distância de Mahalanobis.

Nos dados reais, a combinação que levou aos melhores resultados foi o índice IIV-D com a ferramenta MSV, aumentado a acurácia em 17% com a retenção de 5,49% das variáveis originais. Nos dados simulados, a combinação que levou à melhor acurácia foi o IIV-D com a distância de Mahalanobis, resultando em uma acurácia média de 80,01% com retenção média de 9,35% das variáveis originais. Porém, enfatiza-se que cada combinação mostrou características que podem ser almejadas em determinadas situações, como retenção de menos variáveis, ou análise em bancos de dados altamente correlacionados.

Pesquisas futuras incluem o estudo de técnicas para a minimização do *overffiting*, a combinação de ferramentas em diferentes métodos de seleção de variáveis, como métodos com modelo de busca da forma *filter*, e métodos de seleção de variáveis para classificação em múltiplas classes, onde as bateladas conformes podem ser inseridas em grupos de qualidade.

4.8 Referências

AKAY, M.F. Support vector machines with feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v. 36, n. 2, p. 3240-3247, 2009.

ANZANELLO, M. J. Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes. **Gestão & Produção**, v.16, n.4, p.526-533, 2009.

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. Multicriteria variable selection for classification of production batches. **European Journal of Operational Research**, v. 218, n. 1, p. 97-105, 2012.

ANZANELLO, M.J.; ORTIZ, R.S.; LIMBERGERB, R.P.; MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v.83, n.1, p.209-214, 2013.

BHATTACHARYYA, A. On a measure of divergence between two statistical populations defined by their probability distributions. **Bull. Calcutta Math. Soc.**, v.35, n.4, p.99-109, 1943.

BLOCK, P.C.; PETERSON, E.C.; KRONE, R.; KESLER, K.; HANNAN, E.; O'CONNOR, G.T.; DETRE, KATHERINE. Identification of variables needed to risk adjust outcomes of coronary interventions: Evidence-based guidelines for efficient data collection. **Journal of the American College of Cardiology**, n. 32, v. 1, p. 275-282, 1998.

BLUM, A.L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v.97, n.1, p.245-271, 1997.

BOHLING, G.C.; DAVIS, J.C.; OLEA, R.A.; HARFF, J. Singularity and Nonnormality in the Classification of Compositional Data. **Mathematical Geology**, v. 30, n. 1, p. 5-20, 1998.

BURGES, C.J.C. A tutorial on support vector machines for pattern recognition. **Data mining and Knowledge Discovery**, v.2, n.2, p.121-167, 1998.

CHANG, Y.W.; HSIEH, C.J.; CHANG, K.W.; RINGGAARD, M.; LIN, C.J. Training and testing low-degree polynomial data mappings via linear SVM. **The Journal of Machine Learning Research**, v.99, p.1471-1490, 2010.

CHAOVALITWONGSE, W.A.; FAN, Y.J.; SACHDEO, R.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.

CHEN, H.L.; YANG, B.; LIU, J.; LIU, D.Y. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. **Expert Systems with Applications**, v.38, n.7, p.9014-9022, 2011.

COLEMAN, G. B.; ANDREWS, H. C. Image segmentation by clustering. **Proceedings of the IEEE**, v.67, n.5, p.773-785, 1979.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v.20, n.3 p.273-297, 1995.

COVER, T.; HART, P. Nearest neighbor pattern classification. Information Theory, **IEEE Transactions on Information Theory**, v.13, n.1, p.21-27, 1967.

CRISTIANINI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods**. Cambridge university press, 2000.

DIXON, S.; BRERETON, R.G. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. **Chemometrics and Intelligent Laboratory Systems**, v.95, n.1, p.1-17, 2009.

GARVIN, D.A. **Gerenciando a qualidade: a visão estratégica e competitiva**. Rio de Janeiro, Qualitymark, 2002.

GAUCHI, J.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v.58, n.2, p.171-193, 2001.

GHOSE, A.; IPEIROTIS, P.G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. **Knowledge and Data Engineering, IEEE Transactions on**, v.23, n.10, p.1498-1512, 2011.

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

HAPFELMEIER, A.; ULM, K. A new variable selection approach using Random Forest. **Computational Statistics and Data Analysis**, v.60, p.50-69, 2013.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.; FRANKLIN, J. The elements of statistical learning: data mining, inference and prediction. **The Mathematical Intelligencer**, v.27, n.2, p.83-85, 2005.

HUANG, C.L.; WANG, C.J. A GA-based feature selection and parameters optimization for support vector machines. **Expert Systems with Applications**, v.31, n.2, p. 231-240, 2006.

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

KLUG, M.; MARSHALL, I; VITERBO, E. **Gestão da qualidade**. São Paulo, 2003

KOURTI, T.; MACGREGOR, J.F. Process analysis, monitoring and diagnosis, using multivariate projection methods. **Chemometrics and Intelligent Laboratory Systems**, v.28, n.1, p.3-21, 1995.

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

LONG, X.X.; LI, H.D.; FAN, W.; XU, Q.S.; LIANG, Y.Z. A model population analysis method for variable selection based on mutual information. **Chemometrics and Intelligent Laboratory Systems**. v.121, n.1, p.75-81, 2012.

LUTS, J.; OJEDA, F.; Van de PLAS, R.; De MOOR, B.; Van HUFFEL, S.; SUYKENS, J.A. A tutorial on support vector machine-based methods for classification problems in chemometrics. **Analytica Chimica Acta**, v.665, n.2, p.129-145, 2010.

MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D.L. The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems**, v. 50, n. 1, p. 1–18, 2000.

MAHALANOBIS, P.C. On the generalized distance in statistics. **Proceedings of the National Institute of Sciences (Calcutta)**, v.2, n.1, p.49-55, 1936.

MARTIN, E.B.; MORRIS, A.J.; KIPARISSIDES, C. Manufacturing performance enhancement through multivariate statistical process control. **Annual Reviews in Control**, v.23, n.1, p.35-44, 1999.

MARTINS, S.L.M. **Monitoramento do Controle Estatístico do Processo Utilizando Ferramentas Estatísticas**, 2011. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal de Santa Maria, Brasil.

ROSE-PEHRSSON, S.L.; SHAFFER, R.E.; HART, S.J.; WILLIAMS, F.W.; GOTTUK, D.T.; STREHLEN, B.D.; HILL, S.A. Multi-criteria fire detection systems using a probabilistic neural network. **Sensors and Actuators B: Chemical**, v.69, n.3, p.325-335, 2000.

PEÑA-REYES, C. A.; SIPPER, M. A fuzzy-genetic approach to breast cancer diagnosis. **Artificial Intelligence in Medicine**, v.17, n.2, p.131-155, 1999.

PEREIRA, A.C.; REIS, M.S.; SARAIVA, P.M.; MARQUES, J.C. Madeira wine ageing prediction based on different analytical techniques: UV–vis, GC-MS, HPLC-DAD. **Chemometrics and Intelligent Laboratory Systems**, v.105, n.1, p.43-55, 2011.

POLAT, K.; GÜNEŞ, S. Breast cancer diagnosis using least square support vector machine. **Digital Signal Processing**, v.17, n.4, p.694-701, 2007.

POMPE, B. Measuring statistical dependences in a time series. **Journal of Statistical Physics**, v.73, n.3, p.587-610, 1993.

RACHOW, T; BERGER, S.; BOETTGER, M.K.; SCHULZ, S.; GUINJOAN, S.; YERAGANI, V.; VOSS, A.; BÄR, K.J. Nonlinear relationship between electrodermal activity and heart rate variability in patients with acute schizophrenia. **Psychophysiology**, v.48, n.10, p.1323-1332, 2011.

RAKOTOMAMONJY, A. Variable selection using SVM based criteria. **The Journal of Machine Learning Research**, v.3, n.1, p.1357-1370, 2003.

RODRÍGUEZ-ROSARIO, C.A.; MODI, K.; KUAH, A.M.; SHAJI, A.; SUDARSHAN, E.C.G. Completely positive maps and classical correlations. **Journal of Physics A: Mathematical and Theoretical**, v.41, n.20, 205301, 2008.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, v.23, n.19, p.2507-2517, 2007.

SHANNON, C.E. A mathematical theory of communication. **ACM SIGMOBILE Mobile Computing and Communications Review**, v.5, n.1, p.3-55, 2001.

TIAN, W.M.; HE, Z.; YAN, W. Key Process Variable Identification for Quality Classification Based on PLSR Model and Wrapper Feature Selection. **In Proceedings of 2012 3rd International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012)**, p.263-270. Springer Berlin Heidelberg, 2013.

URTUBIA, A.; PÉREZ-CORREA, J.R.; SOTO, A.; PSZCZOLKOWSKI, P. Using data mining techniques to predict industrial wine problem fermentation. **Food Control**, v.18, n.12, p.1512-1517, 2007.

VAPNIK, V. **The nature of statistical learning theory**. New York: Springer, 1995.

WESTAD, F.; HERSLETH, M.; MARTENS, H. Variable selection in PCA in sensory descriptive and consumer data. **Food Quality and Preference**, v. 14, n. 5, p. 463-472, 2003.

WOLD, S.; SJOSTROM, M.; ERIKSEN, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109-130, 2001.

XIANG, S.; NIE, F.; ZHANG, C. Learning a Mahalanobis distance metric for data clustering and classification. **Pattern Recognition**, v. 41, n. 12, p. 3600-3612, 2008.

5. Considerações finais

Este capítulo apresenta as conclusões a respeito dos estudos realizados nesta dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

O presente trabalho teve como principal objetivo propor métodos de seleção de variáveis com o intuito de classificar bateladas produtivas. A proposição de novos IIVs no primeiro e terceiro artigo permitiu atingir o primeiro objetivo específico desta dissertação. Nos três artigos houve a avaliação de distintas ferramentas de classificação de observações, o segundo objetivo específico desta dissertação. O terceiro objetivo específico foi alcançado nos dois primeiros artigos, comparando os resultados das metodologias propostas a outras metodologias encontradas na literatura. A comparação das variações do método em dados reais e simulados, propostas no terceiro artigo, conduz ao quarto objetivo específico, avaliar a robustez dos métodos em bancos de dados simulados com diferentes níveis de covariância e ruído. Portanto, conclue-se que todos os objetivos específicos foram alcançados, permitindo igualmente afirmar que o objetivo principal deste trabalho foi obtido.

No primeiro artigo foi proposto um novo método para selecionar as variáveis que melhor classificam bateladas produtivas, chamado ROV. Este método consiste em (1) separar os dados históricos em bancos de treino e teste; (2) aplicar a ACP nos dados e gerar o IIV através dos pesos relacionados às variáveis dos componentes retidos, ponderados pela sua variância explicada; (3) classificar as observações através da distância de Mahalanobis, calcular a acurácia da classificação, retirar a variável com menor IIV e repetir a classificação até restar apenas uma variável; (4) definir o subconjunto de variáveis que melhor classifica o processo, verificando a distância dos resultados dos subconjuntos em relação a um hipotético ponto ótimo, e verificar a acurácia do método utilizando as variáveis selecionadas para classificar o banco de testes. Ao ser aplicado em cinco processos industriais, o método obteve uma melhora média de 28,4% na classificação das observações, com a retenção de 8,24% das variáveis. O método mostrou-se eficiente nas diversas proporções de tamanhos dos bancos de treino e teste avaliadas. Quando comparado com o método *Stepwise*, o método ROV produziu resultados superiores e mais confiáveis.

No segundo artigo foi proposto outro método com vistas à classificação de bateladas produtivas, o ROI. Este método consiste em (1) separar os dados históricos em bancos de

treino e teste; (2) aplicar a ACP nos dados e gerar o IIV através dos pesos relacionados às variáveis dos componentes retidos ponderados pela sua variância explicada; (3) classificar as observações do banco de treino utilizando a MSV, calcular a acurácia, retirar a variável com menor IIV e repetir a classificação até restar apenas uma variável; e (4) definir o subconjunto de variáveis que melhor classifica o processo retendo a menor quantidade de variáveis e verificar a acurácia do método utilizando as variáveis selecionadas para classificar o banco de teste. Ao ser aplicado em quatro processos industriais, o método obteve um incremento médio de acurácia de 25,14%, quando comparado à classificação com todas as variáveis, com a retenção média de 7,47% das variáveis originais. Frente ao método OUVV, o ROI mostrou melhor desempenho e menor tempo de processamento.

No terceiro artigo, são utilizadas combinações de IIV, para ordenamento de remoção de variáveis, com ferramentas classificatórias em um método *wrapper* de seleção de variáveis para classificação. Este método consiste em (1) separar os dados históricos em bancos de treino e teste; (2) com as observações do banco de treino, gerar um IIV; (3) classificar as observações do banco de treino, calcular a acurácia e remover a variável com menor IIV, repetindo este processo até restar uma variável; (4) definir o melhor subconjunto de variáveis classificatórias verificando qual está mais próximo de um ponto ótimo hipotético e verificar a acurácia do método utilizando as variáveis selecionadas para classificar as observações do banco de teste. Nas proposições deste artigo são utilizadas duas formas de IIVs; o IIV-D, baseado na distância de Bhattacharyya, e o IIV-IM, baseado na Informação Mútua das variáveis; e três ferramentas classificadoras com características distintas; MSV, KVP e distância de Mahalanobis. Nos dados reais, a combinação que levou aos melhores resultados foi o índice IIV-D com a ferramenta MSV, aumentado a acurácia em 17% com a retenção de 5,49% das variáveis originais. Nos dados simulados, a combinação que levou à melhor acurácia foi o IIV-D com a distância de Mahalanobis, resultando em uma acurácia média de 80,01% com retenção média de 9,35% das variáveis originais. Porém, enfatiza-se que cada combinação mostrou características que podem ser almejadas em determinadas situações, como retenção de menos variáveis, ou análise em bancos de dados altamente correlacionados.

5.2 Sugestões para trabalhos futuros

Como possíveis extensões do estudo apresentado nesta dissertação, sugerem-se as seguintes pesquisas futuras:

a) Extensão dos métodos para o caso de classificação em múltiplas classes;

b) Extensão dos métodos para o caso de múltiplas variáveis de resposta;

c) Extensão dos métodos para o caso de bancos de dados não balanceados;

d) Desenvolvimento de métodos com modelo de busca da forma *filter*, com o intuito de diminuir o *overfitting*; e

e) Desenvolvimento de IIVs para casos onde a importância da variável muda ao longo do processo.