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

TESE DE DOUTORADO

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Porto Alegre, 2017

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito final à obtenção do título de Doutor em Engenharia, na área de concentração em Sistemas de Produção.

Orientador: Professor Michel J. Anzanello, *Ph.D.*

Porto Alegre, 2017

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia 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.

_____

Professor Michel José Anzanello, *Ph.D*.
Orientador PPGEP/UFRGS

_____

Professor Flávio Sanson Fogliatto, *Ph*.D.
Coordenador PPGEP/UFRGS

Banca Examinadora:

Professor Flávio Sanson Fogliatto, *Ph.D.* (PPGEP/UFRGS)

Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

Rafael Scorsatto Ortiz, Dr. (Superintendência Regional do DPF no Rio Grande do Sul)

RESUMO

Nos últimos anos, a espectroscopia no infravermelho (IR) ganhou grande aceitação em diversas áreas de pesquisa por ser uma técnica rápida, simples e não destrutiva que permite a quantificação de diversos componentes químicos em amostras. Apesar de a IR resultar em valores de absorbância que auxiliam na caracterização da amostra, tal técnica acaba por gerar bancos de dados compostos por centenas, ou até milhares, de variáveis altamente correlacionadas e ruidosas, comprometendo o resultado de diversas técnicas de análise multivariada. Dentro deste cenário, esta Tese apresenta novas metodologias para seleção de variáveis, também chamada de seleção de comprimentos de onda quando aplicados em dados de IR, com o intuito de auxiliar o reconhecimento de padrões para o controle de qualidade em diversas áreas. Tais metodologias são apresentadas em três artigos onde as proposições visam à solução de problemas específicos: no primeiro artigo, amostras de erva mate são categorizadas de acordo com seu país de origem através de uma nova metodologia para seleção de variáveis. Para tanto, um problema de Programação Quadrática, combinado com a Informação Mútua entre as variáveis, é utilizado para reduzir a redundância entre as variáveis retidas e maximizar sua relação com o local de origem da amostra; por sua vez, o segundo artigo adequa as proposições do primeiro artigo para um problema de predição, onde o objetivo é determinar a concentração de cocaína e adulterantes em amostras de cocaína laboratoriais e apreendidas; por fim, o terceiro artigo utiliza a estatística do teste de Kolmogorov-Smirnov para duas amostras em uma abordagem de seleção de intervalos de comprimentos de onda com o intuito de identificar falsificações em medicamentos para disfunção erétil. A aplicação dos métodos em bancos de dados com distintas características e a validação dos resultados corrobora a adequabilidade das proposições desta tese.

Palavras-chave: Seleção de Comprimentos de Onda; NIR; FTIR; Classificação; Predição.

ABSTRACT

Over the last few years infrared (IR) spectroscopy gained wide acceptance in many research fields as a quick, simple and non-destructive technique allowing the quantification of many chemical compounds. Although IR provide many absorbance values that helps the sample characterization, this technique also generate databases comprised by hundreds, or even thousands, of highly noisy and correlated wavenumbers, jeopardizing the results of many multivariate analysis techniques. Under such scenario, this thesis presents new variables selection methodologies (also called wavenumber selection when applied in IR data) aimed to recognize patterns for quality control in many areas. Such methodologies are presented in three papers where the propositions are tailored for the solution of specific problems: on the first paper, yerba mate samples are categorized according to their country of origin through a novel variable selection methodology. Thereunto a quadratic programming problem, combined with the Mutual Information among variables, is utilized to reduce the redundancy among variables and increase their relationship with the samples' place of origin; the second paper adequate the first paper propositions for a prediction method which aims to determine cocaine and adulterants concentration in laboratorial and seized cocaine samples; lastly, the third paper uses the two-samples Kolmogorov-Smirnov statistic in an wavenumber interval selection method aimed for the identification of counterfeit erectile dysfunction medicines. The application of the methods in databases with distinct characteristics and the results validation corroborates the suitability of this thesis propositions.

Keywords: Wavenumber selection; NIR; FTIR; Classification; Prediction.

SUMÁRIO

LISTA DE TABELAS

1 Introdução

O rápido avanço de tecnologias para análise e monitoramento de processos e produtos tem gerado volumes crescentes de dados, os quais oferecem oportunidades para a identificação de padrões que expliquem eventos das mais diversas naturezas. Tais dados, no entanto, são tipicamente caracterizados por elevado número de variáveis, o que inviabiliza uma análise minuciosa das mesmas. Além disso, parcela significativa das ferramentas multivariadas de análise perde eficiência frente a dados impregnados por ruído ou multicolinearidade, o que é usualmente percebido em bancos com elevada dimensionalidade (LIU; YU, 2005).

Para quantificar a composição química de produtos, de forma a encontrar padrões que permitam verificar determinadas caraterísticas desejáveis, percebeu-se nos últimos anos um aumento substancial no número de estudos que se apoiam na espectroscopia no infravermelho (IR); tal técnica é tida como de simples execução, rápida e não destrutiva, permitindo estimar a composição química de observações com baixa preparação prévia (CRAIG et al., 2014; LIU; YANG; DENG, 2015; ZHANG; ZHANG; IQBAL, 2013). Apesar de dados do tipo NIR fornecerem diversas informações relevantes para a caracterização de amostras, tipicamente são compostos por diversas características indesejáveis a análises multivariadas. Tal cenário justifica a necessidade da utilização de técnicas de mineração de dados para identificação apropriada de padrões (MAIONE et al., 2016).

A mineração de dados consiste no processo computacional de identificação de padrões em grandes bancos de dados, tendo como principal objetivo extrair informações relevantes e implícitas destes bancos. Dentre as técnicas de mineração de dados, destaca-se a seleção de variáveis (também chamada de seleção de comprimentos de onda quando aplicada a dados do tipo NIR), a qual objetiva identificar as variáveis mais importantes através da remoção de variáveis irrelevantes ou que prejudiquem a interpretação dos dados. Os benefícios desta redução incluem melhor interpretação dos resultados, maior rapidez computacional na geração de modelos e aumento de acurácia de técnicas de predição e classificação. Tais benefícios estão alinhados com as justificativas trazidas pela literatura para seleção de variáveis: (*i*) evitar o *overfitting* de modelos; (*ii*) produzir modelos com menor necessidade de processamento e melhor custo-efetividade; e (*iii*) permitir 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; TIBSHIRANI; FRIEDMAN, 2009; KETTANEH; BERGLUND; WOLD, 2005; SAEYS; INZA; LARRAÑAGA, 2007).

Dentro do escopo desta tese, a seleção de variáveis (ou comprimentos de onda) tem por objetivo criar um modelo de análise selecionando regiões do espectro que sejam significativas, reduzindo a quantidade de variáveis e, consequentemente, removendo dados ruidosos, redundantes, ou irrelevantes. A seleção de regiões relevantes do espectro também contribui na criação de modelos mais simples e, consequentemente, mais fáceis de interpretar, uma vez que tais modelos explicitam não apenas a relação dos comprimentos de onda entre si, como também sua relação com a variável resposta (XIE; YING; YING, 2009; ZHANG; ZHANG; IQBAL, 2013). Por fim, a remoção de comprimentos de onda que não possuem informações relevantes reduz a complexidade do modelo, resultando em ganhos computacionais e de precisão (CHEN et al., 2013).

Existem dois propósitos principais alinhados com a seleção de comprimentos de onda: (i) predição, onde o objetivo é encontrar um conjunto de variáveis independentes que viabilizam melhor predição de variáveis dependente quantitativa (GAUCHI; CHAGNON, 2001; PEREIRA et al., 2011); e (ii) classificação, a qual objetiva encontrar o conjunto de variáveis independentes que melhor insira novas observações em categorias (ANZANELLO et al., 2015; DINIZ et al., 2014). Para atingir tais objetivos, os métodos de seleção de comprimentos de onda se dividem em duas frentes: (i) seleção de comprimentos de onda individuais, como em Anzanello et al. (2015), e (ii) seleção de intervalos de comprimentos de onda, como em Soares et al. (2017) e Marcelo et al. (2014). Os artigos apresentados nesta tese abordam metodologias para classificação e predição voltadas à seleção individual e de intervalos de comprimentos de onda. O primeiro artigo apresenta um método de seleção de comprimentos de onda que visa à identificação do país de origem de amostras de erva mate; por sua vez, o segundo artigo apresenta um método com o intuito de predizer a concentração de cocaína e adulterantes em amostras de cocaína; por fim, um método de seleção de intervalos de comprimentos de onda é proposto no terceiro artigo com o objetivo de identificar falsificações de remédios para disfunção erétil.

1.1 TEMA E OBJETIVOS

O tema da presente tese é a proposição de novas abordagens para seleção de comprimentos de onda com vistas à classificação de amostras e predição de suas propriedades. Os objetivos específicos são:

(i) Criar novos índices de Importância de Comprimentos de onda com vistas a mensurar a relevância das variáveis analisadas;

(ii) Comparar métodos de seleção de intervalos de comprimentos de onda e de seleção individual de comprimentos de onda;

(iii) Comparar os resultados dos métodos propostos a outras metodologias de seleção de variáveis mais difundidas, aplicando-os em bancos de dados reais; e

(iv) Verificar a adequabilidade em dados oriundos de NIR com diferentes características em relação à dimensionalidade e características da variável resposta;

1.2 JUSTIFICATIVA DO TEMA E DOS OBJETIVOS

Nos últimos anos, a espectroscopia no infravermelho (IR) ganhou grande aceitação em diversas áreas de pesquisa por ser uma técnica rápida, simples e não destrutiva que permite a quantificação de diversos componentes químicos em amostras. A IR, combinada com diferentes tipos de técnicas de análise multivariada, tem sido utilizada nas mais diversas áreas de pesquisa, as quais incluem análise forense (BORILLE et al., 2017; MARCELO et al., 2016), engenharia de combustíveis (CRAMER; MORRIS; ROSE-PEHRSSON, 2010; SUN et al., 2011) e engenharia de alimentos (MARQUETTI et al., 2016; ZHANG et al., 2015).

Apesar da IR resultar em valores de absorbância que auxiliam na quantificação de diversos componentes químicos, a técnica acaba por gerar bancos de dados compostos por centenas, ou até milhares, de variáveis altamente correlacionadas e ruidosas, comprometendo o resultado de diversas técnicas de análise multivariada. Dentro deste cenário, a mineração de dados voltada à seleção de regiões relevantes do espectro se mostra necessária tanto para aumentar a qualidade da análise multivariada como para reduzir a influência de dados mal condicionados, gerando assim modelos mais simples e eficientes em termos de interpretação (HE et al., 2014; MARCELO et al., 2014). Tais benefícios justificam as abordagens aqui propostas em termos práticos.

Percebe-se ainda que diversas abordagens clássicas da literatura acabam por não mais produzir modelos satisfatórios quando aplicadas a dados espectrais mais detalhados, os quais são decorrentes da modernização das técnicas experimentais utilizadas na obtenção do NIR. Desta forma, é possível perceber no âmbito acadêmico um grande esforço devotado ao desenvolvimento de abordagens mais robustas e aptas à aplicação em bancos com tendência crescente de dimensionalidade, o que contribui na justificativa acadêmica do tema desta tese.

1.3 ESTRUTURA DA PESQUISA

A pesquisa é realizada em três etapas, onde cada etapa corresponde a um artigo que visa a atender os objetivos específicos supracitados. Com relação à estrutura da tese, cada artigo corresponde a um dos capítulos subsequentes a presente introdução. A Tabela 1-1 apresenta os artigos, ferramentas utilizadas e contribuição científica de cada artigo.

Artigo	Título	Ferramentas utilizadas	Contribuição científica
1	Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin	Programação Quadrática, Informação Mútua, Máquina de Suporte Vetorial, Análise Discriminante, K-vizinhos próximos	Proposição de um novo método de seleção de comprimentos de onda para categorização de amostras de erva mate de acordo com seu país de origem
2	Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples	Programação Quadrática, Informação Mútua, Regressão Linear Múlipla, Regressão por Componentes Principais, Regressão por Mínimos Quadráticos Parciais	Proposição de um novo método de seleção de comprimentos de onda para predição da concentração de cocaína e adulterantes em amostras de cocaína
3	Spectra interval selection to identify counterfeit medicines	Teste de Kolmogorov-Smirnov para duas amostras, Máquina de Suporte Vetorial, Análise Discriminante, K-vizinhos próximos	Proposição de um novo método de seleção de intervalos de comprimentos de onda para categorização de medicamentos falsificados e originais

Tabela 1-1 – Descrição dos artigos da tese

Dentre as principais contribuições desta pesquisa destacam-se: a integração da Programação Quadrática à Informação Mútua voltada à geração de um índice de importância de comprimentos de onda aplicável à seleção de variáveis em problemas de classificação e predição; a proposição de um índice de importância de intervalos de comprimentos de onda através da estatística do teste de Kolmogorov-Smirnov para duas amostras; e a comparação entre a seleção de comprimentos de onda individuais e a seleção de intervalos de comprimentos de onda, duas abordagens utilizadas na literatura.

1.4 DELIMITAÇÕES DA PESQUISA

A pesquisa considera em seu escopo somente ferramentas clássicas de análise multivariada, bancos de dados de NIR voltados a problemas específicos e a validação dos resultados através da comparação com os resultados de técnicas difundidas de seleção de comprimentos de onda ou por especialistas. Desta forma, não foram considerados nesta pesquisa:

- Técnicas de análise multivariada alternativas às existentes na literatura;
- Dados públicos de NIR;
- Modelos alternativos ao *wrapper* com a inclusão de variáveis de forma *forward* ordenada;
- Avaliações de modelos baseados em métricas outras que acurácia e dimensionalidade; e
- A interpretação detalhada dos modelos gerados, analisando apenas os comprimentos de onda selecionados e não suas implicações.

1.5 REFERÊNCIAS

ANZANELLO, M. J., KAHMANN, A., MARCELO, M. C. A., MARIOTTI, K. C., FERRÃO, M. F., ORTIZ, R. S. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, 2015. v. 115, p. 562–569.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, 1997. v. 97, n. 1–2, p. 245–271.

BORILLE, B. T., MARCELO, M. C. A., ORTIZ, R. S., MARIOTTI, K. de C., FERRÃO, M. F., LIMBERGER, R. P. Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2017. v. 173, p. 318–323.

CHEN, M., KHARE, S., HUANG, B., ZHANG, H., LAU, E., FENG, E. Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. **Industrial and Engineering Chemistry Research**, 2013. v. 52, n. 23, p. 7886–7895.

CRAIG, A. P., FRANCA, A. S., OLIVEIRA, L. S., IRUDAYARAJ, J., ILELEJI, K. Application of elastic net and infrared spectroscopy in the discrimination between defective and

non-defective roasted coffees. **Talanta**, 2014. v. 128, p. 393–400.

CRAMER, J. A.; MORRIS, R. E.; ROSE-PEHRSSON, S. L. Use of Genetic Algorithms To Improve Partial Least Squares Fuel Property and Synthetic Fuel Modeling from Near-Infrared Spectra. **ENERGY & FUELS**, 2010. v. 24, p. 5560–5572.

DINIZ, P. H. G. D., GOMES, A. A., PISTONESI, M. F., BAND, B. S. F., de ARAÚJO, M. C. U. Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA. **Food Analytical Methods**, 2014. v. 7, n. 8, p. 1712–1718.

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

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, 2003. v. 3, n. 3, p. 1157–1182.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **Elements**, 2009. v. 1, p. 337–387.

HE, K., CHENG, H., DU, W., QIAN, F. Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. **Chemometrics and Intelligent Laboratory Systems**, 2014. v. 134, p. 79–88.

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

LIU, C.; YANG, S. X.; DENG, L. Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning. **Journal of Food Engineering**, 2015. v. 161, p. 16–23.

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

MAIONE, C., LEMOS, B., DOBAL, A., BARBOSA, F., MELGAÇO, R. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. **Computers and Electronics in Agriculture**, 2016. v. 121, p. 101–107.

MARCELO, M. C. A., MARTINS, C. A., POZEBON, D., FERRÃO, M. F. Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate. **Anal. Methods**, 2014. v. 6, n. 19, p. 7621–7627.

MARCELO, M. C. A., FIORENTIN, T. R., MARIOTTI, K. C., ORTIZ, R. S., LIMBERGER, R.P. Analytical Methods Determination of cocaine and its main adulterants in

seized drugs from Rio Grande do Sul , Brazil , by a Doehlert optimized LC-DAD method. 2016. p. 5212–5217.

MARQUETTI, I., LINK, J. V., LEMES, A. L. G., SCHOLZ, M. B. dos S., VALDERRAMA, P., BONA, E. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. **Computers and Electronics in Agriculture**, 2016. v. 121, p. 313–319.

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**, 2011. v. 105, n. 1, p. 43–55.

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

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, 2017, v.167, n.15, p. 171-178.

SUN, X., ZIMMERMANN, C. M., JACKSON, G. P., BUNKER, C. E., HARRINGTON, P. B. Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography-fast scanning quadrupole ion trap mass spectrometry. **TALANTA**, jan. 2011. v. 83, n. 4, SI, p. 1260–1268.

XIE, L.; YING, Y.; YING, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. **Journal of Food Engineering**, 2009. v. 94, n. 1, p. 34–39.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 128, p. 17–24.

ZHANG, Y., ZHENG, L., LI, M., DENG, X., JI, R. Predicting apple sugar content based on spectral characteristics of apple tree leaf in different phenological phases. **Computers and Electronics in Agriculture**, 2015. v. 112, p. 20–27.

3 Considerações finais

3.1 CONCLUSÕES

A presente tese tem como objetivo principal a proposição de novas metodologias para seleção de comprimentos de onda para aplicação em bancos de dados de espectroscopia no infravermelho oriundos de amostras de naturezas distintas. Nesta tese, as proposições são divididas em três artigos que propões novas abordagens de seleção de comprimentos de onda para a resolução de problemas específicos.

Dentre as principais contribuições do primeiro artigo, destaca-se a proposição de um novo índice de importância de comprimentos de onda, baseado em um problema de programação quadrática composto pelos valores da Informação Mútua entre os pares de comprimentos de onda e entre os comprimentos de onda e a variável resposta. Tal índice é utilizado em uma metodologia de seleção de comprimentos de onda voltada à categorização de amostras de erva mate de acordo com seu país de origem. Através da retenção média de 28% das variáveis originais foi possível categorizar 95,74% das amostras de teste, resultado superior quando comparados a outras metodologias.

Através de uma abordagem multicriterial para seleção de comprimentos de onda, o segundo artigo apresenta uma adaptação da metodologia proposta no primeiro artigo voltada à predição da concentração de cocaína e adulterantes em amostras de cocaína. Para a predição da concentração de cocaína, a regressão por mínimos quadrados parciais apresentou os melhores resultados, tendo um erro médio absoluto de 0,0879 alcançado através da retenção média de 2,03% dos comprimentos de onda originais. Por sua vez, a regressão linear múltipla apresentou os melhores resultados para a predição da concentração de adulterantes. Retendo em média 2,35% dos comprimentos de onda originais esta técnica atingiu um erro médio absoluto de 0,408.

Por fim, o terceiro artigo tem como principal contribuição a apresentação de uma nova metodologia para seleção de intervalos de comprimentos de onda, explorando a comparação de tal abordagem com a seleção de comprimentos de onda individuais. Utilizado para identificar falsificações de remédios para disfunção erétil, o método utiliza a estatística do teste de Kolmogorov-Smirnov para duas amostras para encontrar os intervalos do espectro com maior poder de separação entre as classes “original” e “falsificado”. Entre os bancos de dados analisados, a acurácia nas porções de teste foi de 99,65%, sendo necessária a retenção média de

18,12% do espectro original. Quando comparado à seleção individual de comprimentos de onda, a seleção de intervalos apresentou menor variabilidade dentre as faixas retidas do espectro.

Para atingir o objetivo principal, objetivos específicos foram determinados, os quais foram executados ao longo dos três artigos: dois novos índices de importância de comprimentos de onda foram propostos, indicando o cumprimento do primeiro objetivo específico; o segundo objetivo específico foi atingido no terceiro artigo, onde há a comparação entre métodos de seleção de comprimentos de onda individuais e de seleção de intervalos de comprimentos de onda; já a validação dos resultados dos artigos 1 e 3, através da comparação dos resultados aos resultados de outras metodologias demonstra a consecução do terceiro objetivo específico; por fim, a aplicação dos métodos em bancos de dados com diferentes origens e tipos de variáveis conduz ao quarto e último objetivo específico. Portanto, infere-se que todos os objetivos específicos determinados foram alcançados, permitindo igualmente afirmar que o objetivo principal deste trabalho foi obtido.

3.2 SUGESTÕES PARA TRABALHOS FUTUROS

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

a) Desenvolvimento de novas abordagens de análise multivariada voltadas à seleção de comprimentos de onda;

b) Abordagens para a identificação de observações críticas para a melhora do poder de categorização de amostras; e

c) Desenvolvimento de novos índices de importância voltados à análise de variáveis com diferentes características.