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

TESE DE DOUTORADO

ABORDAGENS MULTIVARIADAS PARA SELEÇÃO DE VARIÁVEIS COM VISTAS À CLASSIFICAÇÃO E PREDIÇÃO DE PROPRIEDADES DE AMOSTRAS

Gabrielli Harumi Yamashita

Porto Alegre, 2021

Gabrielli Harumi Yamashita

ABORDAGENS MULTIVARIADAS PARA SELEÇÃO DE VARIÁVEIS COM VISTAS À CLASSIFICAÇÃO E PREDIÇÃO DE PROPRIEDADES DE AMOSTRAS

Tese 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 Doutor em Engenharia, na área de concentração em Sistemas de Qualidade.

Orientador: Prof. Michel José Anzanello, *Ph.D.*

.

Porto Alegre,

2021

Gabrielli Harumi Yamashita

ABORDAGENS MULTIVARIADAS PARA SELEÇÃO DE VARIÁVEIS COM VISTAS À CLASSIFICAÇÃO E PREDIÇÃO DE PROPRIEDADES DE AMOSTRAS

Banca Examinadora:

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

Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

Filipe Lucini, Dr. (Department of Critical Care Medicine, Cumming School of Medicine/University of Calgary)

AGRADECIMENTOS

Desejo expressar meus sinceros agradecimentos a todos que contribuíram para a realização desta tese. Em especial, agradeço:

À minha mãe, Solange, por me motivar e incentivar durante todos esses anos, pelo suporte e por entender minha ausência devido à distância. Ao meu pai, Fábio, pelo apoio e incentivo. À minha irmã, Flávia, por estar sempre presente, me apoiar e me fazer ser um bom exemplo.

Ao meu marido, José Carlos, pelo companheirismo, paciência, dedicação e por sempre me estimular a dar o meu melhor.

Ao meu orientador e professor, Michel Anzanello, pela orientação, incentivo, dedicação e por todo conhecimento compartilhado.

Aos amigos de longa data, Thiago Bronze, pela amizade, suporte, por estar sempre presente e principalmente por ter me ajudado a entrar no doutorado; e Renata Tilemann, pela amizade, incentivo e principalmente pela diversão!

À Miriam Rocha, pela amizade e parceria, pelas inúmeras horas de brainstorming, pelos ensinamentos e por toda contribuição acadêmica.

Aos amigos de caminhada Felipe Soares e Alessandro Kahmann, e as amigas Érica Roos e Cíntia Franco, pelos momentos de debate, de dúvida e principalmente pelos de descontração.

Ao professor Flávio Fogliatto, pela atenção e aprimoramento da escrita dessa tese.

A todos esses citados, obrigada também pela convivência!

Ao PPGEP-UFRGS pela oportunidade de cursar o mestrado e doutorado

Ao professor Marcelo Farenzena pela disponibilidade e contribuições para a adequação dessa tese.

Ao CNPq, pela bolsa de financiamento sem o qual essa tese não teria sido realizada.

RESUMO

A seleção de variáveis é uma etapa importante para a análise de dados, visto que identifica os subconjuntos de variáveis mais informativas para a construção de modelos precisos de classificação e predição. Além disso, a seleção de variáveis facilita a interpretação e análise dos modelos obtidos, potencialmente reduzindo o tempo computacional de geração dos modelos e o custo/tempo para obtenção das amostras. Neste contexto, a presente tese apresenta proposições inovadoras de abordagens com vistas à seleção de variáveis para classificação e predição de propriedades de amostras de produtos diversos. Tais abordagens são abordadas em três artigos apresentados nesta tese, com intuito de melhorar a precisão dos modelos de classificação e predição em diferentes áreas. No primeiro artigo, integram-se índices de importância de variáveis a sistemáticas de classificação hierárquica para categorizar amostras de espumantes de acordo com seu país de origem. No segundo artigo, para selecionar as variáveis mais informativas para a predição de amostras via PLS, propõe-se um índice de importância de variáveis baseado na Lei de Lambert-Beer combinado a um processo iterativo de seleção do tipo *forward*. Por fim, o terceiro artigo utilizou *cluster* de variáveis espectrais e índice de importância para selecionar as variáveis que produzem modelos de predição mais consistentes. Em todos os artigos dessa tese, os resultados obtidos pelos métodos propostos foram superiores quando comparados a outros métodos tradicionais da literatura voltados à identificação das variáveis mais informativas.

Palavras-chave: Seleção de variáveis. Classificação. Predição. ReliefF. Índice de importância de variáveis. *Cluster* de variáveis. NIR.

ABSTRACT

Variable selection is an important step in data analysis, since it identifies the most informative subsets of variables for build accurate classification and prediction models. In addition, variable selection improves the interpretation and analysis of obtained models, reduces the computational time to build models and reduces the obtained samples costs. In this context, this thesis presents propositions for a variable selection method aiming to classifying and predicting sample properties. Such methods are presented in three papers in this thesis, in order to improve the classification and prediction accuracy in different areas. In first paper, we applied variable importance index coupled with a hierarchical classification technique to identify the country of origin of sparkling wines. In second paper, to select the most informative variables for prediction, a variable improtance index was built based on Lambert-Beer law and an iterative forward process was performed. Finally, in third paper was used clustering of variables and variable importance index to select the variables that produce more consistent prediction models. In all papers of this thesis, when conpared to other traditional methods, our proposition obtained better results.

Key words: Variable selection. Classification. Prediction. RefiefF. Variable importance index. Clustering of variable. NIR.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AAS	*Atomic Absorption Spectrometry*
AG	Algoritmo Genético
AI	*Agreement Index*
biPLS	*backward interval PLS*
Bootstrap-VIP	*Bootstrap combined with Variable Importance on the Projection*
CFS	*Correlation-based Feature Selection*
CW	*Clustering of Wavelengths*
EN-IRRCS	*Elastic Net combined with Iterative Rank PLS Regression Coefficient Screening*
EN-PLSR	*Elastic Net combined with Partial Least Squares Regression*
fiPLS	*forward interval PLS*
FOSS	*Fisher Optimal Subspace Shrinkage*
gPLS	*group PLS*
ICP-MS	*Inductively Coupled Plasma Mass Spectrometry*
ICP-OES	*Inductively Coupled Plasma Optical Emission Spectrometry*
iPLS	*interval PLS*
iRF	*interval Random Frog*
iSPA	*interval Successive Projections Algorithm*
iVISSA	*interval Variable Iterative Space Shrinkage Approach*
IPW-PLS	*Iterative Predictor Weighting PLS*
JMI	*Joint Mutual Information*
KNN	*k-Nearest Neighbor*
LASSO	*Least Absolute Selection and Shrinkage Operator*
LDA	*Linear Discriminant Analysis*
LOD	*Limit of Detection*
LOQ	*Limit of Quantification*
LOO-CV	*Leave-one-out cross validation*
MC-UVE-PLS	*Monte Carlo Uninformative Variables Elimination with PLS*
MRMR	*Minimum Redundancy Maximum Relevance*
MWPLS	*Moving Windows PLS*
NB	*Naive Bayes*

NIR	*Near-infrared spectroscopy*
PCA	*Principal Component Analysis*
PIW	*Potentially Irrelevant Wavelengths*
PLS	*Partial Least Squares*
PRW	*Potentially Relevant Wavelengths*
PSO	*Particle Swarm Optimization*
RMSE	*Root Mean Square Error*
SA	*Simulated Annealing*
siPLS	*synergy interval PLS*
SIS-iPLS	*Independence Screening and interval PLS*
SPA	*Successive Projections Algorithm*
SOM	*Self-Organizing Maps*
SVM	*Support Vector Machine*
UVE-PLS	*Uninformative Variable Elimination in PLS*

SUMÁRIO

1 Introdução

Com o avanço das tecnologias computacionais para coleta e monitoramento de processos e produtos, tem-se observado um rápido crescimento no volume de dados coletados nos mais diversos segmentos e áreas (LIU; YU, 2005). Tais dados geralmente apresentam alta dimensionalidade e são caracterizados por níveis significativos de ruído, sendo compostos por um elevado número de variáveis que, em alguns casos, excedem o número de observações (WANG et al., 2020). Com o aumento da dimensionalidade dos dados, a eficiência de técnicas de aprendizagem, visualização dos dados, compreensão das informações e interpretação das análises são prejudicados devido à presença de variáveis irrelevantes e ruidosas (QU et al., 2019). Assim, a seleção de variáveis desempenha um papel fundamental na análise de grandes volumes de dados, tornando viável a extração de informações relevantes que auxiliam tanto na tomada de decisão, monitoramento de processos e controle de autenticidade (KAMPA et al., 2014).

A seleção de variáveis é o processo de obtenção de um subconjunto específico das variáveis originais de acordo com determinados critérios de seleção (CAI et al., 2018). Dessa forma, a seleção de variáveis procura um subespaço de baixa dimensão para a construção de modelos robustos e confiáveis, eliminando variáveis irrelevantes, ruidosas e inconsistentes sem perder a qualidade das soluções geradas (WANG et al., 2020). Modelos reduzidos não só tendem a aprimorar a interpretação dos fenômenos em análise, mas também podem representar uma redução nos custos associados à coleta e análise de um conjunto maior de variáveis pouco informativas (KAMPA et al., 2014; LEARDI; SEASHOLTZ; PELL, 2002).

Desenvolver métodos apropriados de seleção de variáveis que facilitem a construção de modelos robustos (e a consequente interpretação da relação entre as variáveis e as propriedades de interesse analisadas) tornou-se um desafio para pesquisadores de diversos segmentos (HUANG; XIA, 2017; TONG et al., 2015). Abordagens com vistas à identificação das variáveis mais informativas têm sido aplicadas em diferentes áreas de pesquisas e com distintos propósitos, os quais incluem diagnóstico de transtorno de déficit de atenção e hiperatividade (XIAO et al., 2016), identificação de medicamentos falsos (ANZANELLO et al., 2017), controle de qualidade do diesel (SOARES et al., 2017), monitoramento da qualidade dos processos de fabricação (SHAO et al., 2013), determinação de concentrações de adulterantes em drogas (KAHMANN et

al., 2018), identificação da origem de vinhos (SOARES et al., 2018), e classificação de bateladas produtivas (ANZANELLO; ALBIN; CHAOVALITWONGSE, 2012).

Os métodos de seleção de variáveis podem ser divididos em três abordagens principais, filtro, *wrapper* e *embedded*. Nos métodos do tipo filtro, as variáveis são avaliadas considerando as características de sua natureza (EBRAHIMPOUR; EFTEKHARI, 2018). Essas abordagens, que tipicamente se apoiam em testes estatísticos de significância, são aplicadas previamente a um algoritmo de classificação ou predição (MURSALIN et al., 2017; JIN et al., 2019). Por sua vez, as abordagens do tipo *wrapper* avaliam os subconjuntos de variáveis utilizando algoritmos de aprendizado, de forma a selecionar o subconjunto que apresentar o melhor desempenho; embora avaliem as contribuições de cada variável em instrumentos de predição/classificação, caracterizam-se como um processo mais demorado do que o filtro (BASGALUPP, 2007). Por fim, os métodos *embedded* selecionam o subconjunto de variáveis durante o próprio processo de construção do modelo de classificação, como executado, por exemplo, por algoritmos apoiados em árvores de decisão (RODRIGUEZ-GALIANO et al., 2018).

Dentre as técnicas de seleção do tipo filtro trazidas pela literatura, destacam-se a abordagem Correlation-based feature selection (CFS) (CHORMUNGE; JENA, 2018; MURSALIN et al., 2017; PALMA-MENDOZA et al., 2018), relief e reliefF (JIN et al., 2019; REYES; MORELL; VENTURA, 2015; ZAFRA; PECHENIZKIY; VENTURA, 2012), Minimum Redundancy Maximum Relevance (MRMR) (LI et al., 2017; YAN; JIA, 2019; YUAN et al., 2018), Information Gain (ELMAIZI et al., 2019; JADHAV; HE; JENKINS, 2018; LEFKOVITS; LEFKOVITS, 2017) e Joint Mutual Information (JMI) (BENNASAR; HICKS; SETCHI, 2015; HAN; REN; LIU, 2015). No que diz respeito a técnicas do tipo *wrapper*, destacam-se aquelas apoiadas em Algoritmo Genético (AG) (ANZANELLO et al., 2017; LEARDI, 2000; LEARDI; SEASHOLTZ; PELL, 2002), Particle Swarm Optimization (PSO) (CHEN; ZHOU; YUAN, 2019; QASIM; ALGAMAL, 2018), e Simulated Annealing (SA) (MAFARJA; MIRJALILI, 2017; YAN et al., 2019). Abordagens do tipo *embedded* são exemplificadas por Least Absolute Selection and Shrinkage Operator (LASSO) (LEE; CAI, 2018; ZHANG et al., 2019), e deep learning (LECUN; BENGIO; HINTON, 2015; ZUO et al., 2019).

Embora a literatura reporte vasto repertório de abordagens voltadas à seleção de variáveis, percebe-se que não há um consenso sobre a melhor ou mais eficiente técnica a ser utilizada em cada situação, fazendo com que a identificação das variáveis mais

informativas ainda mostre-se como um tema complexo e aberta a novas abordagens (MUÑOZ-ROMERO et al., 2020). Dessa forma, esta tese propõe o desenvolvimento e aplicação de métodos originais de seleção das variáveis mais relevantes com vistas à classificação e predição de propriedades de amostras. Para tanto, propõe a combinação de técnicas multivariadas existentes, bem como adaptação de alguns de seus mecanismos, voltadas ao aumento da precisão e robustez dos modelos propostos. As técnicas de seleção de variáveis propostas serão combinadas com diferentes ferramentas de classificação e predição, e então comparadas com técnicas reportadas pela literatura para avaliação do seu desempenho e robustez.

1.1 Tema e Objetivos

O tema desta tese consiste em seleção de variáveis com fins de classificação e predição de propriedades de amostras em dados de alta dimensionalidade. O objetivo principal passa pela proposição de novas abordagens de seleção de variáveis em dados sujeitos a elevada dimensionalidade. Como objetivos específicos, lista-se:

(i) Propor e validar um índice de importância baseado nos pesos derivados da técnica ReliefF para selecionar as variáveis mais informativas com vistas à classificação de amostras;

(ii) Avaliar a utilização dos princípios da lei de Lambert- Beer como base para uma nova abordagem de seleção de variáveis;

(iii) Propor uma sistemática de seleção de variáveis em regressão PLS (Partial Least Squares) que se valha da formação de *clusters* de variáveis;

(iv) Avaliar a robustez dos métodos propostos em bancos de dados de diversos processos e áreas frente a outros métodos de seleção de variáveis encontrados na literatura.

1.2 Justificativa do tema e dos objetivos

O rápido avanço de tecnologias voltadas à coleta de dados tem gerado bancos cada vez mais complexos e com alta dimensionalidade (LIU; YU, 2005; CAI et al., 2018). Tais bancos podem conter porções (representadas por variáveis) que não apresentam informações relevantes e que, quando inseridas em modelos preditivos e exploratórios, conduzem a resultados pouco eficientes e conclusões distorcidas (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015). A seleção de variáveis se torna então necessária para reduzir a dimensionalidade dos bancos de dados e melhorar a compreensão do problema, visto que diminui a complexidade do modelo, aumenta a velocidade do aprendizado das ferramentas, aprimora a capacidade de generalização e eleva a precisão em procedimentos de classificação e predição (REMESEIRO; BOLON-CANEDO, 2019).

Além disso, a seleção de variáveis pode reduzir tanto os custos associados à coleta de dados, quanto a demanda de esforço computacional para geração dos modelos (ANZANELLO et al., 2017). A seleção das variáveis mais relevantes através de sistemáticas apoiadas em técnicas multivariadas também reduz a possibilidade de equívocos frente à seleção empírica das variáveis mais informativas, o que habitualmente acontece em cenários práticos. De tal forma, o tema desta tese encontra respaldo prático na potencial redução do volume de dados a serem analisados e na obtenção de resultados mais precisos quando da utilização de modelos mais enxutos.

Em diversas áreas do conhecimento que envolvem grandes volumes de dados, tipicamente observa-se a melhora da qualidade dos resultados obtidos quando técnicas de seleção de variáveis são utilizadas, como na análise de dados espectroscópicos (ANZANELLO et al., 2017), classificação de imagens (ZHU; DORNAIKA; RUICHEK, 2019), reconhecimento facial (VIGNOLO; MILONE; SCHARCANSKI, 2013), classificação de texto (LABANI et al., 2018) e análise de microarranjos de DNA (CHUANG et al., 2011). Diversos segmentos valem-se de tais técnicas, incluindo indústria farmacêutica, forense, alimentícia, química e petroquímica (ANZANELLO et al., 2017; KAHMANN et al., 2018; SOARES et al., 2017, 2018). Esses grandes volumes de dados coletados são provenientes de diversas fontes (sensores, dispositivos móveis, etc.) e apresentam diferentes formatos (discretos, espectroscópicos, texto, imagem, etc.) (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015).

Apesar da literatura apresentar diversas técnicas de seleção de variáveis, entende-se que a natureza dos dados tem grande impacto nos modelos selecionados, fazendo com

que a ênfase das pesquisas para selecionar variáveis relevantes deva se ajustar aos diferentes tipos de conjuntos de dados (CAI et al., 2018). Dessa forma, a realização desta pesquisa encontra justificativa no âmbito teórico através da proposição de novos métodos que permitam a identificação de variáveis relevantes para a construção de modelos robustos de classificação e predição.

1.3 Delineamento do Estudo

Com os objetivos e justificativa desta tese definidos, esta seção apresenta o enquadramento da pesquisa do ponto de vista metodológico, descrevendo o método aplicado para alcançar os objetivos propostos, assim como um resumo das ferramentas utilizadas e contribuições científicas de cada artigo que compõe a tese.

1.3.1 Método de Pesquisa

Do ponto de vista da natureza, esta tese é considerado como pesquisa aplicada, visto que o conteúdo teórico é explorado e direcionado à solução de problemas genéricos (GIL, 2008). A tese se enquadra como pesquisa quantitativa, fazendo uso de análises numéricas e propiciando análises estatísticas da realidade (BERTO; NAKANO, 1999). Quanto aos objetivos, essa pesquisa é classificada como pesquisa exploratória, visto que permite conhecer e ter uma visão geral do problema, possibilitando a construção de hipóteses para solucioná-lo (GIL, 2008).

1.3.2 Método de Trabalho

A presente tese é composta por três etapas, cada uma delas correspondendo a um artigo com o intuito de atender os objetivos da tese. No primeiro artigo é proposto um método que integra os pesos gerados pelo reliefF à classificação hierárquica com vistas a identificar as variáveis mais relevantes para a classificação de amostras de espumantes de acordo com seu país de origem. A classificação hierárquica apoiou-se na estratégia “um contra o resto” para determinar as classes que compõem cada nível da hierarquia. Nessa estratégia, o banco de dados apresentando *n* classes é dividido em *n* sub-problemas binários, procedendo-se então à classificação de uma classe contra todas as remanescentes. Em cada nível da classificação hierárquica realizou-se um procedimento

iterativo de seleção de variáveis, identificando-se o subconjunto de variáveis que melhor distingue as classes das amostras de acordo com sua procedência. Para isso, um índice de importância de variáveis que orienta a eliminação sistemática de variáveis do tipo *backward* é construído com base nos pesos calculados pelo algoritmo reliefF.

No segundo artigo, um método para selecionar variáveis que melhorem a qualidade das predições de propriedades de amostras químicas é proposto e operacionalizado em duas fases. Na primeira fase, um novo método de divisão de subconjuntos de variáveis é proposto. Para tanto, as amostras do banco de dados são separadas em quartis de acordo com a variável de resposta (absorbância), e então em cada variável é calculada a distância entre a média da absorbância das amostras pertencentes ao primeiro e quarto quartis. Em seguida, as variáveis são divididas em subconjuntos de acordo com a distância associada a cada uma. Assim, variáveis que possuem as maiores distâncias são agrupadas no mesmo subconjunto e assim sucessivamente. Na segunda fase é realizado o processo iterativo de seleção de subconjuntos de variáveis do tipo *forward*. Esse procedimento promove uma inserção ordenada dos subconjuntos de variáveis no modelo de regressão Partial Least Squares (PLS), iniciando pelo subconjunto que contém as variáveis com as maiores distâncias. Os subconjuntos responsáveis pelos menores desvio de predição são selecionados.

No terceiro artigo propõe-se a clusterização de variáveis espectrais com o intuito de selecionar as variáveis mais relevantes para predição de propriedades de produtos. Para tanto, inicialmente as variáveis são clusterizadas via k-means utilizando a correlação como medida de agrupamento. Em seguida, as variáveis inseridas em cada *cluster* são avaliadas de acordo com seu desempenho de predição em um modelo PLS. *Clusters* cujas variáveis apresentam erro de predição menor do que o erro de predição obtido pelo modelo composto por todas as variáveis são pré-selecionados. *Clusters* cujas variáveis não satisfazem essa condição são então submetidos a uma nova análise via regressão LASSO (Least Absolute Selection and Shrinkage Operator), a qual avalia a existência de variáveis específicas em termos de informação em tais *clusters*. Se as variáveis selecionadas pelo LASSO reduzirem o erro (gerando resíduos menores do que os obtidos com todas as variáveis), as mesmas são pré-selecionadas; caso contrário, todas as variáveis inseridas naquele *cluster* são descartadas. As variáveis pré-selecionadas pelas etapas acima são então submetidas a nova rodada de seleção com base na sua relevância medida por quatro índices de importância de variáveis para calibração de modelos PLS.

Na Tabela 1.1 são apresentados os três artigos que compõem a tese, as ferramentas utilizadas e contribuições científicas de cada artigo. Ressalta-se que os artigos estão apresentados no formato de submissão ou publicação dos periódicos internacionais.

Tabela 1. 1. Descrição dos artigos da tese.

Estudos	Título	Ferramentas utilizadas	Contribuição científica
Artigo 1 [(a)]	Hierarchical classification of sparkling wine samples according to the country of origin based on the most informative chemical elements	ReliefF, Classificação Hierárquica, k-Nearest Neighbor(KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA)	a) Proposição de um método de seleção de variáveis para a classificação de amostras de espumantes de acordo com seu país de origem; b) Utilização do reliefF como índice de importância de variáveis; c) Utilização da classificação hierárquica para classificar bancos de dados com mais de 2 classes.
Artigo 2 [(b)]	Selecting relevant wavelength intervals for PLS calibration based on wavelength importance ranking	Divisão de intervalos, Partial Least Square Regression (PLS)	a) Nova abordagem de divisão de intervalos para a seleção de variáveis espectroscópicas.
Artigo 3 [(c)]	Wavelength clustering and selection as a preliminary step for PLS calibration	*Cluster* de variáveis, Índice de importância de variáveis, Least absolute shrinkage and selection operator (LASSO), Partial Least Square Regression (PLS)	a) Utilização de *clusters* de variáveis em conjunto com índice de importância de variáveis para selecionar as variáveis mais informativas para previsão via PLS.

(a) Artigo publicado no periódico Food Control
(b) Artigo submetido ao periódico Journal of Chemometrics
(c) Artigo a ser submetido ao periódico Chemometrics and Intelligent Laboratory Systems

1.4 Delimitações do Estudo

A presente pesquisa tem foco no desenvolvimento e aplicação de métodos de seleção de variáveis, estando limitado à utilização de ferramentas e conceitos existentes na literatura. As abordagens de seleção de variáveis utilizadas nessa tese são do tipo *wrapper* e *embedded*, não sendo utilizadas abordagens do tipo filtro.

Para a classificação das amostras, os modelos KNN foram construídos utilizando apenas a distância Euclidiana e o SVM com kernel linear. Por sua vez, para a predição de propriedades de amostras avaliou-se somente o desempenho da PLS, não tendo sido consideradas modelagens via PLS não-linear.

Em relação à abrangência, a pesquisa concentra-se nas áreas alimentícia, petroquímica, biologia e farmacêutica. Quanto ao banco de dados, foram analisados apenas bancos de dados supervisionados, com fins de classificação e predição. Com isso, os desempenhos das abordagens propostas se restringem a métricas de avaliação como acurácia, raiz do erro quadrático médio (RMSE) e coeficiente de determinação R^2. Por fim, não foram avaliados aspectos de redução de custos decorrentes da menor necessidade de coleta de dados.

1.5 Referências

ANZANELLO, M. J. et al. A genetic algorithm-based framework for wavelength selection on sample categorization. **Drug Testing and Analysis**, 2017.

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, abr. 2012.

BASGALUPP, M. P. **Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócio**. [s.l.] PUC-RS, 2007.

BENNASAR, M.; HICKS, Y.; SETCHI, R. Feature selection using Joint Mutual Information Maximisation. **Expert Systems with Applications**, v. 42, n. 22, p. 8520–8532, dez. 2015.

BERTO, R. M. V. S.; NAKANO, D. N. A produção científica nos anais do encontro nacional de engenharia de produção: um levantamento de métodos e tipos de pesquisa. **Production**, v. 9, n. 2, p. 65–75, dez. 1999.

BOLÓN-CANEDO, V.; SÁNCHEZ-MARROÑO, N.; ALONSO-BETANZOS, A. Recent advances and emerging challenges of feature selection in the context of big data. **Knowledge-Based Systems**, v. 86, p. 33–45, 2015.

CAI, J. et al. Feature selection in machine learning: A new perspective. **Neurocomputing**, v. 300, p. 70–79, jul. 2018.

CHEN, K.; ZHOU, F.-Y.; YUAN, X.-F. Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. **Expert Systems with Applications**, v. 128, p. 140–156, ago. 2019.

CHORMUNGE, S.; JENA, S. Correlation based feature selection with clustering for high dimensional data. **Journal of Electrical Systems and Information Technology**, v. 5, n. 3, p. 542–549, dez. 2018.

CHUANG, L.-Y. et al. A hybrid feature selection method for DNA microarray data. **Computers in Biology and Medicine**, v. 41, n. 4, p. 228–237, abr. 2011.

EBRAHIMPOUR, M. K.; EFTEKHARI, M. Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 173, p. 51–64, fev. 2018.

ELMAIZI, A. et al. A novel information gain based approach for classification and dimensionality reduction of hyperspectral images. **Procedia Computer Science**, v. 148, p. 126–134, 2019.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6ª ed. São Paulo: Atlas S. A., 2008.

HAN, M.; REN, W.; LIU, X. Joint mutual information-based input variable selection for multivariate time series modeling. **Engineering Applications of Artificial Intelligence**, v. 37, p. 250–257, jan. 2015.

HUANG, X.; XIA, L. Improved kernel PLS combined with wavelength variable importance for near infrared spectral analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 168, n. June, p. 107–113, 2017.

JADHAV, S.; HE, H.; JENKINS, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. **Applied Soft Computing**, v. 69, p. 541–553, ago. 2018.

JIN, L. et al. A ReliefF-SVM-based method for marking dopamine-based disease characteristics: A study on SWEDD and Parkinson's disease. **Behavioural Brain Research**, v. 356, n. September 2018, p. 400–407, jan. 2019.

KAHMANN, A. et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018.

KAMPA, K. et al. Sparse optimization in feature selection: application in neuroimaging. **Journal of Global Optimization**, v. 59, n. 2–3, p. 439–457, 8 jul. 2014.

LABANI, M. et al. A novel multivariate filter method for feature selection in text classification problems. **Engineering Applications of Artificial Intelligence**, v. 70, p. 25–37, abr. 2018.

LEARDI, R. Application of genetic algorithm-PLS for feature selection in spectral data sets. **Journal of Chemometrics**, v. 14, n. 5–6, p. 643–655, 2000.

LEARDI, R.; SEASHOLTZ, M. B.; PELL, R. J. Variable selection for multivariate calibration using a genetic algorithm: Prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. **Analytica Chimica Acta**, v. 461, n. 2, p. 189–200, 2002.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 27 maio 2015.

LEE, C.-Y.; CAI, J.-Y. LASSO variable selection in data envelopment analysis with small datasets. **Omega**, dez. 2018.

LEFKOVITS, S.; LEFKOVITS, L. Gabor Feature Selection Based on Information Gain. **Procedia Engineering**, v. 181, p. 892–898, 2017.

LI, Y. et al. A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection. **Mechanical Systems and Signal Processing**, v. 91, p. 295–312, jul. 2017.

LIU, H.; YU, L. Toward Integrating Feature Selection Algorithms for Classification and Clustering. **Knowledge Creation Diffusion Utilization**, v. 17, n. 4, p. 491–502, 2005.

MAFARJA, M. M.; MIRJALILI, S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection. **Neurocomputing**, v. 260, p. 302–312, out. 2017.

MUÑOZ-ROMERO, S. et al. Informative variable identifier: Expanding interpretability in feature selection. **Pattern Recognition**, v. 98, p. 107077, fev. 2020.

MURSALIN, M. et al. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. **Neurocomputing**, v. 241, p. 204–214, jun. 2017.

PALMA-MENDOZA, R.-J. et al. Distributed correlation-based feature selection in spark. **Information Sciences**, nov. 2018.

QASIM, O. S.; ALGAMAL, Z. Y. Feature selection using particle swarm optimization-based logistic regression model. **Chemometrics and Intelligent Laboratory Systems**, v. 182, p. 41–46, nov. 2018.

QU, Y. et al. Non-unique decision differential entropy-based feature selection. **Neurocomputing**, jul. 2019.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in Biology and Medicine**, v. 112, p. 103375, set. 2019.

REYES, O.; MORELL, C.; VENTURA, S. Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context. **Neurocomputing**, v. 161, p. 168–182, ago. 2015.

RODRIGUEZ-GALIANO, V. F. et al. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. **Science of The Total Environment**, v. 624, p. 661–672, maio 2018.

SHAO, C. et al. Feature selection for manufacturing process monitoring using cross-validation. **Journal of Manufacturing Systems**, v. 32, n. 4, p. 550–555, out. 2013.

SOARES, F. et al. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, n. June, p. 171–178, 2017.

SOARES, F. et al. Element selection and concentration analysis for classifying South America wine samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 150, n. March, p. 33–40, 2018.

TONG, P. et al. Improvement of NIR model by fractional order Savitzky–Golay derivation (FOSGD) coupled with wavelength selection. **Chemometrics and Intelligent Laboratory Systems**, v. 143, p. 40–48, abr. 2015.

VIGNOLO, L. D.; MILONE, D. H.; SCHARCANSKI, J. Feature selection for face recognition based on multi-objective evolutionary wrappers. **Expert Systems with Applications**, v. 40, n. 13, p. 5077–5084, out. 2013.

WANG, S. et al. Structured learning for unsupervised feature selection with high-order matrix factorization. **Expert Systems with Applications**, v. 140, p. 112878, fev. 2020.

XIAO, C. et al. An integrated feature ranking and selection framework for ADHD characterization. **Brain Informatics**, v. 3, n. 3, p. 145–155, 2 set. 2016.

YAN, C. et al. Hybrid binary Coral Reefs Optimization algorithm with Simulated Annealing for Feature Selection in high-dimensional biomedical datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 184, p. 102–111, jan. 2019.

YAN, X.; JIA, M. Intelligent fault diagnosis of rotating machinery using improved multiscale dispersion entropy and mRMR feature selection. **Knowledge-Based Systems**, v. 163, p. 450–471, jan. 2019.

YUAN, F. et al. Data mining of the cancer-related lncRNAs GO terms and KEGG pathways by using mRMR method. **Mathematical Biosciences**, v. 304, p. 1–8, out. 2018.

ZAFRA, A.; PECHENIZKIY, M.; VENTURA, S. ReliefF-MI: An extension of ReliefF to multiple instance learning. **Neurocomputing**, v. 75, n. 1, p. 210–218, jan. 2012.

ZHANG, R. et al. A variable informative criterion based on weighted voting strategy combined with LASSO for variable selection in multivariate calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 184, p. 132–141, jan. 2019.

ZHU, R.; DORNAIKA, F.; RUICHEK, Y. Learning a discriminant graph-based embedding with feature selection for image categorization. **Neural Networks**, v. 111, p. 35–46, mar. 2019.

ZUO, R. et al. Deep learning and its application in geochemical mapping. **Earth-Science Reviews**, v. 192, p. 1–14, maio 2019.

5. CONSIDERAÇÕES FINAIS

Este capítulo apresenta as conclusões finais sobre a pesquisa descrita nesta tese, salientando pontos sobre os objetivos, métodos e resultados obtidos; seguida pelas indicações de pesquisas futuras.

5.1 Conclusões

A presente tese teve como objetivo principal a proposição de novas abordagens para a seleção de variáveis com vistas à classificação de amostras em categorias (as quais podem estar associadas a níveis de qualidade ou autenticidade), bem como e predição de propriedades de produtos. Neste trabalho foram apresentados três artigos finalizados de modo a alcançar os objetivos específicos declarados. São eles: (*i*) propor e validar um índice de importância baseado nos pesos derivados da técnica ReliefF para selecionar as variáveis mais informativas com vistas à classificação de amostras; (*ii*) avaliar a utilização dos princípios da lei de Lambert-Beer como base para uma nova abordagem de seleção de variáveis; (*iii*) propor uma sistemática de seleção de variáveis em regressão PLS que se valha da formação de *clusters* de variáveis; e (*iv*) Avaliar a robustez dos métodos propostos em bancos de dados de diversos processos e áreas frente a outros métodos de seleção de variáveis encontrados na literatura.

Os objetivos (*i*) e (*iv*) foram atingidos no primeiro artigo, o qual apresentou a classificação hierárquica como solução para a classificação de amostras de bancos de dados com mais de duas classes; o método também utiliza os pesos resultantes do algoritmo reliefF para construir um índice de importância de variável. Tal índice é utilizado como base para a remoção ordenada de variáveis através de um procedimento iterativo do tipo *backward* e assim selecionar as variáveis mais relevantes para a classificação de amostras de espumantes de acordo com seu país de origem.

O Artigo 1 teve como principal motivação prática contribuir para o controle de autenticidade das bebidas em relação à sua origem, visto que é um assunto de grande importância para os consumidores e produtores, pois garante a qualidade de fabricação do produto e tem impacto na sua correta precificação e reputação. Neste estudo, foram analisadas 111 amostras de espumantes procedentes de quatro países: Argentina, Brasil, França e Espanha. Como resultado, obteve-se 100% de acurácia, ou seja, todas as

amostras foram classificadas corretamente quanto ao seu país de origem, utilizando 3 das 12 variáveis disponíveis no banco de dados (Magnésio, Lítio e Potássio). O magnésio foi o elemento que melhor discriminou as amostras de espumantes brasileiros, sendo um metal encontrado em altas concentrações no solo do Brasil e podendo estar associado ao uso de produtos fitossanitários nas plantações de uva. Por sua vez, a Argentina foi melhor classificada utilizando o lítio, fato que se justifica pelos grandes depósitos desse elemento em seu território; e o potássio, que pode estar relacionado ao uso de fertilizantes no solo. Por fim, a alta concentração de lítio nas amostras da Espanha fez com que esse elemento fosse selecionado para distinguir as amostras da França e Espanha. O resultado obtido pelo método proposto nesse artigo superou outros métodos encontrados na literatura. Até onde se tem conhecimento, este foi o primeiro estudo a utilizar a classificação hierárquica para resolver o problema da classificação com mais de duas classes em dados de concentração de elementos químicos. Além disso, também foi o primeiro estudo a utilizar o índice de importância de variáveis baseado nos pesos gerados pelo reliefF como ferramenta de seleção de variáveis aplicado em cada nível da classificação hierárquica

Os objetivos (*ii*) e (*iv*) foram alcançados no segundo artigo, que propôs uma abordagem utilizando a lei de Lambert-Beer como suporte para a seleção dos subconjuntos de variáveis mais relevantes para a predição de propriedades das amostras. O método foi dividido em duas etapas. Na primeira etapa as amostras são divididas em quartis de acordo com a variável resposta e então é construído um perfil de distância utilizando a distância entre a média da absorbância do primeiro e quarto quartil para cada variável. Presume-se que as variáveis com grandes distâncias entre os quartis são responsáveis pelas variações na variável resposta e, assim, consideradas mais informativas para a predição. Na sequência o perfil de distância é dividido em subconjuntos de variáveis de acordo com a distância obtida por cada variável. Na segunda etapa é realizado o processo iterativo de inserção ordenada dos subconjuntos de variáveis do tipo *forward*, iniciando pelo subconjunto que contém as variáveis com as maiores distâncias entre o primeiro e quarto quartis, e assim selecionar as variáveis mais relevantes para a predição de propriedades dos produtos via PLS.

O método proposto no Artigo 2 foi aplicado a três bancos de dados espectroscópicos de domínio público, com um total de 10 variáveis de resposta. Sabe-se que os dados espectroscópicos apresentam alta dimensionalidade e são compostos por variáveis altamente correlacionadas e ruidosas. Por isso, é importante selecionar as

variáveis mais informativas para a construção de modelos mais precisos para a predição. Os resultados obtidos neste artigo foram comparados com a predição utilizando todas as variáveis originais e a quatro métodos de seleção de variáveis tradicionais na literatura. Quando aplicado no conjunto de treino, o método proposto alcançou o menor erro ($RMSE_{tr}$) e maior coeficiente de determinação (R^2_{tr}) em 7 das 10 variáveis respostas, e o menor erro de predição ($RMSE_{ts}$) e maior índice de concordância (AI_{ts}) em 6 variáveis respostas no conjunto de teste. Por fim, o método se mostrou significativamente melhor que os métodos aos quais foi comparado.

As proposições do Artigo 2 contribuíram na construção de modelos mais precisos de predição, sugerindo o método como competitivo frente a abordagens conhecidas da literatura. Com base na literatura pesquisada, este foi o primeiro estudo a propor a divisão do banco de dados em quartis referentes à variável de resposta e então identificar variáveis relevantes para a predição de acordo com a distância entre a média da absorbância do primeiro e quarto quartis de cada variável.

Por fim, os objetivos (*iii*) e (*iv*) foram cumpridos no terceiro artigo, o qual propôs um novo método de seleção de variáveis utilizando *cluster* de variáveis, LASSO e índice de importância de variáveis. No método proposto, inicialmente é realizada a clusterização das variáveis via k-means, utilizando a correlação entre as variáveis como medida de agrupamento. Em seguida, a habilidade preditiva das variáveis pertencentes a cada *cluster* é avaliada utilizando a regressão PLS, sendo o erro de predição calculado. As variáveis pertencentes aos *clusters* que apresentam erro inferior ao obtido utilizando todas as variáveis são pré-selecionadas. Caso contrário, é aplicado o algoritmo LASSO para identificar se há variáveis informativas nesses *clusters* que não obtiveram bom desempenho. Com isso, se as variáveis selecionadas pelo LASSO apresentarem erro inferior ao obtido com todas as variáveis, elas também são pré-selecionadas. Por fim, as variáveis pré-selecionadas pelos *clusters* e, se houver, as variáveis pré-selecionadas pelo LASSO, são ordenadas em termos de importância por meio de quatro diferentes índices de importância de variáveis. Na última etapa, cada índice de importância de variáveis apoia um processo iterativo de remoção das variáveis não informativas do tipo *backward*, acabando por selecionar as variáveis mais relevantes para a predição.

O método proposto no Artigo 3 foi avaliado utilizando 12 bancos de dados espectroscópicos de domínio público referente a diferentes áreas de estudo como bebidas, alimentos, diesel, agricultura e farmácia. Os resultados obtidos foram apresentados de

duas formas. A primeira, sem a análise do LASSO, as variáveis dos *clusters* que apresentam o erro maior do que o erro com todas as variáveis são excluídas e os índices de importância de variáveis são construídos apenas com as variáveis pré-selecionadas pelos *clusters*. A segunda utiliza todas as etapas do método (*cluster* de variáveis, LASSO e índice de importância). Os resultados sugerem que em 9 dos 12 bancos de dado não seria necessário a análise pelo LASSO, pois as variáveis selecionadas pelo LASSO apresentam o erro maior do que o erro com todas as variáveis, logo sua pré-seleção não é justificada. Nesses 9 bancos de dados, apenas as variáveis pré-selecionadas pelos *clusters* foram analisadas. Nos demais 3 bancos, a utilização das variáveis selecionadas pelo LASSO melhorou o desempenho da predição, tendo em vista que as variáveis selecionadas dos *clusters* hipoteticamente menos relevantes contribuíram com a geração de modelos PLS mais precisos. Assim, percebe-se que não é recomendado descartar inicialmente os *clusters* que não obtiveram bom desempenho, e sim aplicar análises adicionais (LASSO, no caso deste estudo) para identificar se há alguma variável informativa alocada àqueles *clusters*. Em relação aos índices de importância de variáveis, os melhores resultados foram alcançados com os índices baseados nos parâmetros do PLS.

Quando comparado a métodos baseados nos modelos PLS tradicionais da literatura, o método proposto obteve os melhores resultados no conjunto de treino em todos os bancos de dados e em 7 dos 12 bancos no conjunto de teste. Além disso, percebe-se que os índices de importância de variáveis obtiveram melhores resultados quando aplicados ao banco de dados pré-selecionado do que quando aplicados ao banco de dados completo. Como contribuições deste artigo, destaca-se a utilização de um algoritmo do tipo *wrapper* para identificar as variáveis mais relevantes entre as pré-selecionadas pelos *clusters*, e assim eliminar as variáveis redundantes. Além disso, até onde se tem conhecimento, é o primeiro estudo a integrar *cluster* de variáveis e índice de importância de variáveis para a seleção de variáveis.

Por fim, com base no que foi exposto acima, conclui-se que esta tese cumpriu todos os objetivos específicos propostos e contribuiu para o avanço dos estudos na área de seleção de variáveis com fins de classificação e predição de propriedades das amostras.

5.2 Sugestões para trabalhos futuros

Como possíveis extensões da pesquisa apresentada nesta tese, sugerem-se as seguintes ações para pesquisas futuras:

a) Analisar diferentes métodos de seleção de variáveis para, em conjunto com a classificação hierárquica, obter mais precisão da classificação em bancos de dados com mais de duas classes;
b) Refinar os intervalos propostos pelo Artigo 2, avaliando a correlação e redundância das variáveis inseridas nos intervalos;
c) Utilizar testes estatísticos para identificar as variáveis que serão excluídas dos clusters no método proposto pelo Artigo 3;
d) Utilizar diferentes algoritmos para clusterização de variáveis como o Generative Topographic Mapping (GTM);
e) Utilizar o método proposto no Artigo 3 para fins de classificação.