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

Gabrielli Harumi Yamashita

ABORDAGENS MULTIVARIADAS PARA A SELEÇÃO DE VARIÁVEIS COM VISTAS À CARACTERIZAÇÃO DE MEDICAMENTOS

Porto Alegre
2015

Gabrielli Harumi Yamashita

Abordagens multivariadas para a seleção de variáveis com vistas à caracterização de medicamentos

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 José Anzanello, *Ph.D.*

Porto Alegre

2015

Gabrielli Harumi Yamashita

Abordagens multivariadas para a seleção de variáveis com vistas à caracterização de medicamentos

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 José 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. (DEQUI/UFRGS)

Professora Márcia Elisa Soares Echeveste, Dr. (PPGEP/UFRGS)

YAMASHITA, Gabrielli Harumi. ***Abordagens multivariadas para a seleção de variáveis com vistas à caracterização de medicamentos***, 2015. Dissertação (Mestrado em Engenharia) – Universidade Ferderal do Rio Grande do Sul, Brasil.

RESUMO

A averiguação da autenticidade de medicamentos tem se apoiado na análise de perfil por espectroscopia de infravermelho (ATR-FTIR). Contudo, tal análise tipicamente gera dados caracterizados por elevado número de variáveis (comprimentos de onda) ruidosas e correlacionadas, necessitando assim da aplicação de técnicas para seleção das variáveis mais relevantes e informativas, tornando os modelos preditivos e exploratórios mais robustos. Esta dissertação testa sistemáticas para a seleção de variáveis com vistas à clusterização e classificação de medicamentos. Para tanto, inicialmente faz-se uso dos parâmetros oriundos da Análise de Componentes Principais (ACP) para a geração de três índices de importância de variáveis; tais índices guiam um processo iterativo de eliminação de variáveis com vistas a uma clusterização mais consistente, medida através do *Silhouette Index*. Na sequência, utiliza-se o Algoritmo Genético (AG) combinado com a ferramenta de classificação *k nearest neighbor* (*k*NN) para selecionar o subconjunto de variáveis que resultem na maior acurácia média com propósito de classificação das amostras em dois grupos, originais ou falsificados. Por fim, aplica-se a divisão dos dados ATR-FTIR em intervalos para selecionar as regiões espectroscópicas mais relevantes para a classificação das amostras via *k*NN; na sequência, aplica-se o AG para refinar os intervalos retidos anteriormente. A aplicação dos métodos de seleção de variáveis propostos permitiu realizar clusterizações e classificações mais precisas com base em um subconjunto reduzido de variáveis.

Palavras-chave: Seleção de variáveis, Clusterização, Análise de Componentes Principais, Algoritmo Genético, Classificação, Seleção por Intervalos.

YAMASHITA, Gabrielli Harumi. ***Multivariate approaches to variable selection in order to characterize medicines***, 2015. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

ABSTRACT

The investigation of the authenticity of drugs has relied on the profile analysis by infrared spectroscopy (ATR-FTIR). However, such analysis typically yields a large number of correlated and noisy variables (wavelengths), which require the application of techniques for selecting the most informative and relevant variables to improve model ability. This thesis test an approach to variable selection aimed at clustering and classifying drug samples. For that matter, it derives three variable importance indices based on Principal Component Analysis (PCA) components that guide an iterative process of variable elimination; clustering performance based on the reduced sets is assessed via Silhouette Index. Next, we combine the Genetic Algorithm (GA) with the *k* nearest neighbor classification technique (kNN) to select the subset of variables yielding the highest average accuracy for classifying samples into authentic or counterfeit categories. Finally, we split the ATR-FTIR data into intervals to select the most relevant spectroscopic regions for sample classification via kNN; we then apply GA to refine the ranges previously retained. The implementation of the proposed variable selection methods led to more accurate clustering and classification procedures based on a small subset of variables.

Keywords: Variable selection, clustering, Principal Component Analysis, Genetic Algorithm, classification, interval selection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACP	Análise de Componentes Principais
AG	Algoritmo Genético
ATR-FTIR	*Attenuated Total Reflectance – Fourier Transform Infrared*
*k*NN	*k Nearest Neighbor*
PSO	*Particle Swarm Optimization*
PLS	*Partial Least Squares*
SI	*Silhouette Index*
SVM	*Suport Vector Machines*

LISTA DE SÍMBOLOS

i	Amostra
SI_i	*Silhouette Index* relativo a cada amostra
a(i)	Média das distâncias de uma amostra a todas as outras pertencentes ao mesmo *cluster*
b(i)	Média das distâncias de uma amostra a todas as outras pertencentes ao *cluster* mais próximo
j	Variável
a	Componente principal
w_{ja}	Pesos de cada variável relacionada ao componente principal
λ_a	Variância explicada de cada componente principal
n	Número do índice
v_{jn}	Índice de importância de variavel

SUMÁRIO

1 Introdução

1.1 Considerações Iniciais

A comercialização de medicamentos falsificados oferece sérios riscos à saúde pública, pois o uso de medicamentos não confiáveis em termos de composição, condições de manipulação e qualidade da matéria-prima pode levar à falha do tratamento e até à morte (GAUDIANO *et al*., 2007; WHO, 2012). O crescimento das falsificações é resultante da extrema dificuldade em rastrear os canais de fabricação e circulação das falsificações, facilidade do acesso dos falsificadores às tecnologias para imitar medicamentos originais, fiscalização inadequada e facilidade de compra desses produtos pela internet (WHO, 2012; FERNANDEZ *et al*., 2011; SACRÉ *et al*., 2010).

A gravidade do uso desses medicamentos fraudulentos motiva a intensificação de pesquisas voltadas à identificação das falsificações (LOPES e WOLFF, 2009). Dentro deste contexto, a análise de perfil por espectroscopia de infravermelho ATR-FTIR (*Attenuated Total Reflectance - Fourier Transform Infrared*) tem sido amplamente utilizada por gerar resultados confiáveis de forma rápida e sem a necessidade de um pré-tratamento das amostras (ORTIZ *et al*., 2013). Contudo, a análise ATR-FTIR gera resultados caracterizados por elevado número de variáveis ruidosas e correlacionadas, evidenciando a importância da aplicação de técnicas que permitam remover as variáveis não informativas e garantir a construção de modelos consistentes de classificação ou predição (ANZANELLO *et al*., 2013).

A seleção de variáveis é importante para aprimorar a interpretabilidade do modelo proposto, bem como para gerar modelos mais robustos e confiáveis (LEARDI *et al*., 2002). O objetivo principal da seleção de variáveis é identificar o subconjunto que possui as informações mais significativas para a construção dos modelos. Tal busca deve ser considerada um quesito imprescindível, visto que, quando se realiza uma predição ou classificação com a totalidade das variáveis originais, pode-se incorrer em resultados inapropriados por conta de variáveis ruidosas, correlacionadas e irrelevantes (RAYMER *et al*., 2000; POON *et al*., 2013).

Esta dissertação é composta por três artigos abordando a seleção de variáveis com propósito de clusterização e classificação de amostras de medicamentos. No primeiro artigo é proposto e testado um método de seleção de variáveis a partir da eliminação *backward* de

variáveis, as quais são ordenadas através de 3 Índices de Importância de Variáveis apoiados nos parâmetros as Analise de Componentes Principais (ACP); as amostras são então classificadas através da técnica de classificação *k Nearest Neighbor* (*k*NN). O segundo artigo aborda a aplicação do Algoritmo Genético (AG) para selecionar subconjuntos de variáveis e então classificar as amostras em duas classes. O terceiro artigo apresenta a divisão do espectro em intervalos seguida pela classificação via *k*NN em cada intervalo; após selecionadas as regiões mais relevantes, aplica-se o AG com vistas à eliminação das variáveis irrelevantes contidas nos intervalos anteriormente selecionados.

1.2 Objetivos

O objetivo principal da dissertação é testar sistemáticas de seleção de variáveis com vistas à clusterização e classificação de amostras de medicamentos.

Os seguintes objetivos específicos são apresentados:

- Selecionar as variáveis ATR-FTIR mais relevantes para clusterização de amostras de medicamentos através de um índice de importância gerado a partir de parâmetros oriundos da ACP;
- Criar Índices de Importância de Variáveis apoiados na ACP que conduzam a uma remoção ordenada de variáveis;
- Aplicar o AG para selecionar os subconjuntos de variáveis ATR-FTIR mais relevantes para a classificação de amostras de medicamentos;
- Utilizar a técnica de divisão do espectro em intervalos para identificar as regiões que contenham as variáveis mais relevantes para classificação de amostras de medicamentos em duas classes.

1.3 Justificativa do Tema e dos Objetivos

O avanço de tecnologias de análises baseadas em espectro infravermelho tornou possível extrair informações precisas acerca da composição de amostras de maneira fácil e rápida. Entretanto, o grande volume de dados tipicamente gerados por tais técnicas pode

conter partes que não possuem informações relevantes, causando distorções aos modelos preditivos e exploratórios gerados e conclusões errôneas. Consequentemente, o tratamento dos dados obtidos passou a exigir modelos mais complexos, que vão além da tradicional calibração univariada (COSTA FILHO e POPPI, 2002; ANZANELLO *et al*., 2013). Desta forma, a sitemática aqui proposta encontra respaldo prático na potencial redução do volume de dados a serem analisados quando da caracterização de amostras de medicamentos.

No contexto acadêmico, percebe-se um interesse no desenvolvimento de abordagens para selecionar variáveis mais relevantes no contexto de ATR-FTIR, garantindo a predominância daquelas que irão definir uma estrutura consistente dos modelos e tornando mais confiável a informação oriunda dos resultados obtidos (MEHMOOD *et al.*, 2012). Assim, a realização desta pesquisa se justifica em função do estudo de possibilidades de combinação de técnicas multivariadas que permitam a identificação de variáveis espectroscópicas relevantes para compor modelos robustos de clusterização e classificação.

1.4 Procedimentos Metodológicos

Quanto aos objetivos, essa dissertação é classificada como pesquisa exploratória, dado que permite conhecer o problema e possibilita construir hipóteses para solucioná-lo. Quanto à natureza, é considerada como pesquisa aplicada, tendo em vista que seu conteúdo teórico é explorado e direcionado à solução de problemas genéricos (GIL, 2002). A dissertação é enquadrada como pesquisa quantitativa, pois faz uso de análises numéricas.

No primeiro artigo o método exposto para selecionar variáveis relevantes para a clusterização de amostras de medicamentos é sequenciado em quatro passos. Inicialmente são coletados dados procedentes da análise espectroscópica ATR-FTIR, em seguida é aplicado a ACP nos dados e, com os parâmetros gerados pela ACP, elaborado três índices de importância de variáveis. No terceiro passo é ralizada a eliminação ordenada das variáveis, onde a variável com menor índice de importância é retirada e uma nova clusterização é realizada até restar apenas uma variável. Por fim é construído um gráfico relacionando o valor de SI médio com as variáveis retidas, de forma a visualizar o conjunto de variáveis que obteve o maior valor de SI médio.

Para classificar amostras de medicamentos em duas classes, originais e falsificados, o segundo artigo apresenta um método dividido em quatro passos. No primeiro passo são coletados dados espectroscópicos gerados pela análise ATR-FTIR. Posteriormente são definidos os parâmetros e critérios para a execução do AG, e então é aplicado o AG nos bancos de dados de forma a gerar subconjuntos de variáveis que são utilizadas na classificação das amostras via *k*NN. A classificação é repetida diversas vezes seguidas com cada subconjunto, obtendo uma acurácia média, e o subconjunto a ser retido pelo método é aquele que apresentar o maior valor de acurácia média. O método é finalizado com a construção de um gráfico relacionando a acurácia média com o número de variáveis retidas.

No terceiro artigo o método aplicado para selecionar variáveis que melhorem a qualidade da classificação de medicamentos é dividido em duas fases. Na primeira fase é utilizada a técnica de divisão do espectro em intervalos equidistantes, onde as variáveis contidas em cada intervalo são então utilizadas em uma ferramenta de classificação *k*NN. Em seguida, a acurácia média resultante da classificação é calculada para cada intervalo, permitindo identificar os intervalos que contêm as variáveis mais informativas (são mantidas aquelas regiões cuja acurácia utilizando intervalos do espectro é maior do que a acurácia utilizando todo o espectro). Na segunda fase é aplicado o AG nos intervalos selecionados na primeira fase e nas combinações dos intervalos que não se sobrepõem, de modo a excluir as variáveis ruidosas remanescentes.

1.5 Estrutura da Dissertação

A dissertação está organizada em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos e as justificativas, bem como o método de pesquisa adotado. A delimitação e estrutura do trabalho completam o capítulo.

O segundo capítulo apresenta o primeiro artigo, que propõe a criação de três índices de importância de variáveis apoiados nos parâmetros oriundos da ACP; cada índice conduzirá a uma eliminação sistemática de variáveis do tipo *backward* e, após cada remoção, as amostras são clusterizadas através do algoritmo *k-means.* O método proposto pretende selecionar o subconjunto de variáveis responsáveis pelo agrupamento mais consistente das amostras em dois grupos: medicamentos originais ou falsos.

O terceiro capítulo apresenta o segundo artigo, que testa a utilização do algoritmo genético para selecionar os subconjuntos de variáveis mais relevantes para a classificação das amostras via *k*NN. O método tem seu desempenho avaliado através da acurácia média, que indica a precisão das classificações realizadas; o subconjunto de variáveis a ser retido é aquele que apresentar a maior acurácia média.

O quarto capítulo traz o terceiro artigo, o qual divide o método em duas fases. Na primeira fase é feita a divisão do espectro em regiões e em cada uma delas realizada a classificação via *k*NN, podendo-se assim identificar as regiões que contêm as variáveis mais relevantes. Na segunda fase, aplica-se o AG combinado com o *k*NN para remover as variáveis ruidosas das regiões selecionadas e obter o subconjunto de variáveis mais relevantes para a classificação das amostras de medicamentos.

O quinto e último capítulo traz a conclusão do trabalho, na qual são avaliados os principais resultados frente aos objetivos almejados e as delimitações citadas. Essa seção traz ainda sugestões para desdobramentos futuros.

1.6 Delimitações do Estudo

Constituem-se em restrições do presente estudo:

- O trabalho não irá desenvolver uma nova ferramenta ou algoritmo para seleção de variáveis, clusterização e classificação, utilizando ferramentas já existentes e combinando-as para obter o resultado esperado;
- As variáveis são selecionadas com objetivo de clusterização e classificação, e não de predição; e
- O banco de dados estudado se restringe a amostras de apenas dois medicamentos, Cialis® e Viagra®.

1.7 Referências Bibliográficas

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGER, 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, p. 209-214, 2013.

COSTA FILHO, P. A.; POPPI, R. J. *Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose*. Química Nova, v. 25, p. 46-52, 2002.

FERNANDEZ, F. M.; HOSTETLER, D.; POWELL, K.; KAUR, H.; GREEN, M.; MILDENHALL, D. C.; NEWTON, P. N. *Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries*. Analyst, v. 136, p. 3073-3082, 2011.

GAUDIANO, M.C.; DI MAGGIO, A.; ANTONIELLA, E.; VALVO, L.; BERTOCCHI, P.; MANNA, L.; BARTOLOMEI, M.; ALIMONTI, S.; RODOMONTE, A. L. *An LC method for the simultaneous screening of some common counterfeit and sub-standard antibiotics Validation and uncertainty estimation.* Journal of Pharmaceutical and Biomedical Analysis, v. 48, p. 303-309, 2008.

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, p. 189-200, 2002.

LOPES, M. B.; WOLFF, J-C. *Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging*. Analytica Chimica Acta, v. 633, p. 149-155, 2009.

POON, L. K. K.; ZHANG, N. L.; LIU, A. H. *Model-based clustering of high-dimensional data: Variable selection versus facet determination*. International Journal of Approximate Reasoning, v. 54, p. 196-215, 2013.

RAYMER, M. L.; PUNCH, W. F.; GOODMAN, E. D.; KUHN, L. A.; JAIN, A. K. *Dimensionality Reduction Using Genetic Algorithms*. IEEE Transaction on Evolutionary Computation, v. 4, p. 164-171, 2000.

SACRÉ, P.Y.; DECONINCK, E.; DE BEER, T.; COURSELLE, P.; VANCAUWEBERGHE, R.; CHIAP, P.; CROMMEN, J.; DE BEER, J. O. *Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines*. Journal of Pharmaceutical and Biomedical Analysis, v. 53, p. 445-453, 2010.

WHO Media centre, *Medicines: spurious/falsely-labelled/ falsified/counterfeit (SFFC) medicines.* (Fact sheet nº 275), 2012. Disponível em < http://www.who.int/mediacentre/factsheets/fs275/en/index.html>. Acesso em 04 de março de 2015.

2 Primeiro Artigo: Sistemática para seleção de variáveis do tipo ATR-FTIR para clusterização de medicamentos

Gabrielli Harumi Yamashita

Michel José Anzanello

Resumo

A análise de perfil por espectroscopia de infravermelho ATR-FTIR tem sido vastamente utilizada na identificação de medicamentos falsificados. Os dados gerados, no entanto, tipicamente contêm um elevado número de variáveis ruidosas e correlacionadas. Este artigo aborda um método de seleção de variáveis para a clusterização de amostras de medicamentos em dois grupos, originais ou falsificados. Três índices de importância de variáveis são elaborados com base nos parâmetros gerados da aplicação da análise de componentes principais nos dados ATR-FTIR. Cada índice guiará a eliminação de variáveis através de uma sistemática do tipo *backward*; após cada variável removida, as amostras são agrupadas via *k-means* e a qualidade do agrupamento avaliada através do *Silhouette Index* (SI). Os índices que se apoiam no parâmetro variância explicada obtiveram resultados mais expressivos quando aplicados ao banco de dados do Viagra® e Cialis®, apresentando maiores SI médios e retendo um menor número de variáveis.

Palavras-chave: ATR-FTIR, Seleção de variáveis, Índice de importância de variáveis, clusterização, *Silhouette Index*.

2.1 Introdução

A venda de medicamentos falsificados vem aumentando significativamente em todo o mundo, acarretando sérios riscos para a saúde pública (GAUDIANO *et al*., 2007). O crescimento do mercado desses medicamentos se deve à facilidade do acesso às tecnologias necessárias para copiar os medicamentos originais e em adquirir produtos pela internet sem receita médica e de sites fraudulentos (FERNANDEZ *et al*., 2011; SACRÉ *et al*., 2010). A

World Health Organization (WHO, 2012) define que os medicamentos falsificados são aqueles fraudulentamente rotulados quanto à identidade ou fonte, e cita que o uso desses medicamentos pode prejudicar um tratamento e até mesmo resultar em morte. Isso se dá, segundo Anzanello *et al*. (2013), pela ausência de informações seguras sobre a produção do medicamento, bem como sobre os ingredientes farmacológicos ativos e a origem das matérias-primas utilizadas.

A maioria dos medicamentos, desde os mais complexos para tratar casos de risco de morte até os mais comuns como analgésicos, apresenta versões falsificadas (WHO, 2012). No mercado brasileiro, o banco de dados da polícia federal mostrou que, no período de janeiro de 2007 a setembro de 2010, 80% dos medicamentos falsificados apreendidos eram amostras de Cialis® e Viagra® (JUNG *et al*,. 2012). Esses medicamentos têm um alto índice de falsificação por conter inibidores da fosfodiesterase tipo 5 (PDE-5) para disfunção erétil, são caracterizados por elevado custo, fazem sucesso no mercado e são amplamente vendidos pela internet devido ao constragimento dos consumidores (HOLZGRABE *et al.*, 2011).

Para verificar a autenticidade dos medicamentos, diversas abordagens laboratoriais utilizam a ferramenta de perfil por espectroscopia de infravermelho ATR-FTIR (*Attenuated Total Reflectance - Fourier Transform Infrared*), que gera resultados confiáveis de forma rápida (ORTIZ *et al*., 2013). No entanto, dados espectroscópicos apresentam número de variáveis substancialmente maior que a quantidade de amostras, fazendo com que a aplicação de técnicas para seleção de variáveis mostre-se importante na identificação de variáveis relevantes que permitam remover as variáveis não informativas e facilitar a interpretação dos resultados gerados pelos modelos (ANZANELLO *et al*., 2013; XIAOBO *et al*., 2010).

Dentre as técnicas de análise multivariadas utilizadas para interpretar dados gerados pelo ATR-FTIR destacam-se Análise de Componentes Principais (ACP) e as técnicas de clusterização. A ACP reduz as variáveis originais através de combinações lineares das mesmas, dando origem aos componentes principais (SACRÉ *et al*., 2010). A clusterização, por sua vez, busca formar grupos de amostras de medicamentos (objeto de estudo deste artigo) com base nas similaridades e diferenças de tais amostras. A seleção de variáveis, por sua vez, é um passo importante para garantir uma clusterização consistente, pois o uso de variáveis irrelevantes e ruidosas na formação dos *clusters* pode prejudicar a qualidade do agrupamento (ANZANELLO e FOGLIATTO, 2011). Estudar a clusterização de

medicamentos falsificados é importante no cenário forense, pois pode ajudar na identificação de fontes falsificadoras, auxiliando as operações de combate às falsificações (LOPES e WOLFF, 2009).

Neste contexto, este artigo apresenta um método para selecionar as variáveis ATR-FTIR mais relevantes para clusterização de amostras em dois grupos (alusivos a amostras autênticas ou falsificadas). Para tanto, utiliza-se a ACP como base para a elaboração de três índices de importância de variáveis. Cada índice, integrado a um processo de eliminação de variáveis do tipo *backward*, conduz ao conjunto de variáveis mais relevantes para a formação de agrupamentos consistentes. Por fim, os três índices serão comparados quanto ao seu desempenho em termos da qualidade da clusterização obtida.

Este artigo está estruturado como segue. Na seção 2 será apresentado um referencial teórico sobre ACP e clusterização, em seguida, na seção 3, será descrito o método aplicado na pesquisa. Na seção 4, são apresentados os resultados obtidos em conjunto com as discussões referentes. Por fim, a seção 5 traz as considerações finais.

2.2 Referencial Teórico

2.2.1 Ferramentas multivariadas

Os métodos de análise de variáveis são divididos em dois grupos, a estatística univariada, que trata as variáveis de maneira isolada e a estatística multivariada que analisa as variáveis de forma conjunta (VICINI e SOUZA, 2005). De acordo com Rencher (2002), a análise multivariada consiste em um conjunto de métodos e técnicas utilizados quando se obtém um grande número de variáveis resultantes da repetição de medições feitas em amostras e é preciso a interpretação teórica dos dados obtidos. Vicini e Souza (2005) destacam que uma grande quantidade de informação deve ser processada antes de ser transformada em conhecimento e com isso cresce a necessidade da utilização de ferramentas estatísticas que apresentem uma visão global do fenômeno.

As técnicas de análise multivariada têm o objetivo de simplificar o conjunto de dados, explicando a maior parte da variabilidade do sistema em um conjunto reduzido de dimensões (RENCHER, 2002). Tais técnicas são recomendadas para cenários em que as variáveis

estudadas estejam correlacionadas, assim seus efeitos não podem ser significamente interpretados separadamente, fazendo com que a técnica desembarace a sobreposição de informações fornecidas pela correlação e elucide informações relevantes (RENCHER, 2002; HAIR JR., 2005). Dentre tais técnicas, destacam-se a Análise de Componentes Principais e a Análise de Cluster, cujos fundamentos são apresentados na sequência.

A Análise de Componentes Principais (ACP) é, segundo Jolliffe (2002), uma das mais antigas e conhecidas técnicas multivariadas. Para Kim *et al.* (2002) e Sebzalli e Wang (2001), a ACP tem como ideia central a redução de dimensionalidade de um conjunto de dados formado por elevado número de variáveis correlacionadas, tentando explicar a maior parte da variabilidade desse conjunto.

A ACP transforma as variáveis do banco de dados original em um conjunto menor de novas variáveis, chamadas de componentes principais, que se baseiam em combinações lineares das variáveis originais (ou seja, transforma as variáveis originais em um novo sistema de coordenadas ortogonais) (SEBZALLI e WANG, 2001; ANDERSON, 2003).

Rencher (2002) reforça que o primeiro componente principal é uma combinação linear que retrata a maior variância do sistema, ou seja, ele representa a dimensão onde as observações apresentam uma maior variabilidade; o segundo componente principal procura uma dimensão ortogonal ao primeiro e que tenha a maior variabilidade subsequente e assim por diante. Assim, segundo Jolliffe (2002), a ACP consegue diminuir a quantidade de dimensões estudadas e não perder informação relevante, pois a maior variabilidade está sendo representada pelos componentes principais.

Por sua vez, a análise de *cluster* (clusterização) é uma técnica multivariada que procura padrões em um conjunto de dados observados e tenta encontrar um agrupamento onde as observações dentro de um *cluster* sejam as mais semelhantes possíveis e diferentes das observações dos demais *clusters* (RENCHER, 2002).

De acordo com Hair *et al.* (1995), existem dois tipos de algoritmo para a clusterização dos dados, hierárquicos e não hierárquicos. Downs e Barnard (2002) dizem que a análise hierárquica constrói um dendograma ou um diagrama de árvore para representar a hierarquia construída entre os indivíduos, onde cada nível do dendograma representa uma partição do conjunto de dados; com base nisso é possível definir o número de *clusters* e a pertinência de

cada observação a cada grupo. Os algoritmos não hierárquicos, segundo Hair *et al.* (1995), agrupam as observações em uma quantidade *k* de *clusters* definida pelo pesquisador. O método mais utilizado é o *k-means* que, após a definição de *k*, calcula um centroide para cada grupo e utiliza a distância euclidiana das observações aos *k* centróides para alocar as observações.

Para avaliar se a clusterização teve um bom desempenho, utiliza-se o *Silhouette Index* (SI), que mede a semelhança de uma observação em relação às outras observações que estão em seu *cluster* e as observações pertencentes aos *clusters* vizinhos (KAUFMAN e ROUSSEEUW, 2005). Conforme Anzanello e Fogliatto (2011) a cada observação *i* está associado um SI_i, que varia de -1 a +1. Valores de SI_i próximos a -1 indicam que as observações provavelmente foram inseridas erroneamente no *cluster*, quando próximos a zero indicam que as observações não estão claramente definidas quanto ao *cluster* pertencente e valores próximos a +1 indicam que a observação foi alocada corretamente ao *cluster*. A relação do SI_i é apresentada na Eq. (1).

$$SI_i = \frac{b(i)-a(i)}{\max\{b(i);\, a(i)\}} \qquad (1)$$

onde $a(i)$ é a média das distâncias da *i*-ésima observação a todas as outras pertencentes ao mesmo *cluster*, $b(i)$ é a média das distâncias dessa *i*-ésima observação a todas as outras alocadas no *cluster* mais próximo.

2.2.2 Abordagens para seleção de variáveis com vistas à clusterização

A seleção de variáveis relevantes é um quesito importante na análise de clusterização, visto que, quando se conduz o agrupamento com a totalidade das variáveis originais, pode-se incorrer em inserções inapropriadas de observações em *clusters* (grupos) por influência de variáveis ruidosas e correlacionadas (POON *et al.*, 2013).

A fim de minimizar a influência de tais variáveis ruidosas (que não definem a estrutura do *cluster*) é recomendada a redução das mesmas, as quais são selencionadas variáveis através de métodos de projeção, seleção de variáveis ou uma combinação de ambos (BRUSCO, 2004; MEHMOOD *et al.*, 2012).

No intuito de selecionar variáveis que resultem em um melhor desempenho na análise de clusterização, diversos autores utilizam o método da ACP. Anzanello *et al.* (2013) aplicam a ACP em um banco de dados de amostras de medicamentos para disfunção erétil. Os parâmetros oriundos da ACP permitem a geração de um índice de importância de variáveis, o qual é integrado ao método *k*NN (*k*-nearest neighbor) para a classificação das amostras em medicamentos originais ou falsificados; os resultados mostram que com menos de 2% das variáveis originais, pode-se obter um bom desempenho na classificação. Com propósitos semelhantes, Lopes e Wolff (2009) selecionam as variáveis mais relevantes com vistas à classificação de amostras de um medicamento antibiótico em original ou falsificado; objetivam ainda agrupar essas amostras por nível de semelhança das variáveis. Para isso, utilizam a ACP para reduzir a dimensionalidade do conjunto de dados obtidos e geram gráficos dos componentes principais de forma a visualizar indícios de agrupamento, os quais confirmados com a ferramenta *k-means*. Os autores concluem que a seleção de variáveis por ACP se mostra eficiente na classificação dos medicamentos e na clusterização das amostras falsificadas.

Ding e He (2004) destacam que os componentes principais obtidos da ACP são a solução contínua para os indicadores de grupos gerados na clusterização pelo método *k-means*; eles aplicam sua teoria em um banco de dados de DNA e de artigos da *internet* e concluem que a combinação de ACP com o *k-means* é eficiente para selecionar variáveis e agrupá-las. Por sua vez, Urtubia *et al.* (2007) estudam o comportamento da fermentação de vinhos; para reduzir a dimensionalidade do conjunto de dados, aplicou-se a ACP, selecionando as variáveis representativas da maior variabilidade e então agrupando as amostras via *k-means*. Os resultados permitiram formar *clusters* de observações com problemas semelhantes na fermentação e identificar a causa desse problema.

As técnicas de ACP na seleção de variáveis para clusterização podem ser vistas ainda nos estudos de Latifoğlu *et al.* (2008) para agrupar indivíduos saudáveis ou que possuem aterosclerose, enquanto que Xue *et al.* (2011) aplicam a ACP em dados de geoquímica marinha para selecionar as variáveis e depois clusterizar de acordo com a ordem de degradação. Por fim Yücel e Sultanoğlu (2012) utilizam a ACP na mineração das variáveis obtidas de amostras de composições químicas de mel de diferentes procedências; a clusterização buscou evidenciar semelhanças entre as propriedades de cada localidade.

2.2.2.1 Índice de importância de variáveis

Em procedimentos de seleção de variáveis para clusterização, diversos autores atribuem pesos às variáveis de acordo com os parâmetros oriundos da ACP, como visto em Honda *et al.* (2009), Brusco e Cradit (2001) e Steinley e Brusco (2008). Tais pesos medem a importância das variáveis e a intensidade com que elas afetam o processo de clusterização.

Anzanello *et al.* (2013) propõem um índice de importância de variáveis utilizando os pesos fornecidos pela aplicação da ACP no conjunto de dados. Assim, as variáveis com maiores pesos são as que explicam uma maior variabilidade do sistema e têm uma maior importância na clusterização. Com o índice de importância gerado, eliminam-se iterativamente as variáveis de menor peso e verifica-se o desempenho da clusterização após cada eliminação; o subconjunto conduzindo aos agrupamentos mais consistentes é recomendado.

Diante de diversos métodos para atribuir pesos às variáveis, Gnanadesikan *et al.* (1995) realizaram uma comparação entre nove métodos de geração de pesos para selecionar as variáveis que incorram em uma melhor qualidade na clusterização; os autores concluíram que a geração de pesos através de estimativas da variabilidade dentro dos *clusters* e entre os *clusters* são geralmente mais eficazes quando comparados aos métodos existentes de normalização. Visto isso, Huang *et al.* (2005) propõem um novo tipo de algoritmo, o W-*k*-*means*, que adiciona ao método *k-means* uma etapa de geração de pesos com base nas distâncias das variáveis dentro do *cluster* e entre os *clusters* vizinhos. O algoritmo demonstrou que se tem um melhor desempenho na clusterização quando as variáveis são selecionadas pelos pesos atribuidos.

A atribuição de pesos de forma a criar uma hierarquia de importância entre as variáveis são apresentados também nos estudos de Makarenkov e Legendre (2001) que, baseados no algoritmo proposto por De Soete (1986), atribuem pesos às variáveis de modo a produzir distâncias euclidianas adequadas para a representação da estrutura dos dados; por fim, utiliza a simulação de Monte Carlo para identificar as situações onde o algoritmo de geração de pesos é uma vantagem para o desempenho da clusterização. Já Modha e Spangler (2002) utilizam uma generalização da função discriminante de Fischer para gerar pesos às variáveis de modo a conduzir um agrupamento que minimize a dispersão dentro do *cluster* e

maximize a dispersão entre os *clusters*. Por fim, Xu *et al.* (2007) combinam um algoritmo de otimização PSO (*Particle Swarm Optimization*) à regressão PLS (*Partial Least Squares*) para ponderar as variáveis importantes e selecioná-las de modo a excluir as variáveis ruidosas que contenham informações irrelevantes.

2.3 Método

Ao analisar dados espectroscópicos é notório observar que o número de variáveis é significativamente maior que o número de amostras (ANZANELLO *et al*., 2013). Com isso, recomenda-se o uso da ACP para selecionar as variáveis mais relevantes, pois, segundo Ortiz *et al*. (2013), esta é a técnica que tem sido aplicada com maior sucesso na interpretação dos dados ATR-FTIR. Anzanello e Fogliatto (2011) destacam que para garantir uma clusterização consistente, a seleção de variáveis é um passo importante, visto que a utilização de variáveis irrelevantes pode prejudicar a qualidade dos agrupamentos.

O método aqui apresentado para selecionar as variáveis mais relevantes para clusterização de amostras de medicamentos é dividido em quatro passos: (1) coletar dados FTIR para análise, (2) aplicar ACP nos dados e gerar um índice de importância de variáveis, (3) eliminar as variáveis irrelevantes através de um procedimento iterativo (a variável com menor índice de importância é retirada e uma nova clusterização é realizada até que reste apenas uma variável), e (4) construir perfil gráfico relacionando SI médio *versus* número de variáveis. Os passos (2), (3), e (4) serão repetidos para os três índices de importância de variáveis propostos. Esses passos serão detalhados a seguir.

2.3.1 Passo 1 - Coletar dados FTIR para análise

Inicialmente, são coletados dados oriundos da análise de perfil por espectroscopia de infravermelho ATR-FTIR em amostras de medicamentos. De acordo com Anzanello *et* al. (2013), a análise por ATR-FTIR tem sido vastamente adotada como uma ferramenta para caracterizar medicamentos falsificados. Ortiz *et al.* (2013) destacam que tal análise gera resultados confiáveis de forma rápida e tornou-se popular por dispensar o pastilhamento ou outro tratamento prévio da amostra, sendo necessário apenas colocar a amostra em contato direto com a superfície do cristal do aparelho para a obtenção de dados espectroscópicos.

Na espectroscopia no infravermelho por Transformada de Fourier (FTIR) a radiação é guiada na amostra através do interferômetro de Michelson, que consiste em dividir um feixe de luz em dois caminhos (diferentes comprimentos de onda), refleti-los de volta e recombiná-los em um anteparo, produzindo um padrão de interferência e resultando em um sinal de interferograma. Ao realizar a transformada de Fourier no sinal obtêm-se um espectro de frequências idêntico ao da espectroscopia convencional, porém no FTIR a medida de um único espectro é bem mais rápida porque as informações de todas as freqüências são colhidas simultaneamente (MEDEIROS, 2009). A maioria dos espectrômetros de infravermelhos modernos pode ser convertido para caracterizar amostras através da reflexão total atenuada (ATR) que utiliza uma propriedade de reflexão interna total, na qual uma amostra é colocada em contato com um elemento de alto índice de refração; a radiação atravessa o elemento de reflexão sendo refletida e direcionada para um detector (HIND *et al*., 2001).

Os espectros gerados pela análise ATR-FTIR tem como característica serem formados por bandas contínuas em vez de respostas discretas, porém essas faixas contínuas são compostas por muitas respostas discretas em estreita proximidade de comprimento de onda, que são identificadas como as variáveis a serem estudadas (LEARDI *et al*.,2002).

2.3.2 Passo 2 - Aplicar ACP nos dados e gerar índices de importância de variáveis

Aplica-se a ACP nos dados coletados no passo 1, de forma a obter parâmetros relevantes sobre a variabilidade dos dados. Os parâmetros de interesse são os pesos (w_{ja}) de cada variável *j* relacionada ao componente principal *a*, fornecido pelos autovetores da matriz de correlação das variáveis; e a variância (λ_a) explicada por cada componente principal *a*, obtidos através dos autovalores de cada componente retido.

Com base em tais parâmetros, elabora-se três índices de importância de variáveis (v_{jn}), n = 1, 2, 3; apresentados nas Eqs. (2) a (4). Nestes índices, as variáveis com os maiores pesos geradas por componentes principais retidos que explicam grande parte da variabilidade apresentarão altos valores de v_{jn} (DUDA *et al*., 2001), denotando variáveis mais relevantes com vistas à estruturação dos agrupamentos.

$$v_{j1} = \sum_{a=1}^{A} \left| w_{ja} \right| \tag{2}$$

$$v_{j2} = \sum_{a=1}^{A} \left| w_{ja} \right| \cdot \lambda_a \tag{3}$$

$$v_{j3} = \sum_{a=1}^{A} \left| w_{ja} \right|^{\lambda_a} \quad (4)$$

2.3.3 Passo 3 - Eliminar as variáveis irrelevantes e agrupar amostras

Nesta etapa, as variáveis são ordenadas de forma decrescente de acordo com seu índice de importância, o que orientará a eliminação sistemática (*backward)* das variáveis tidas como menos relevantes (menores v_j). O procedimento iterativo de eliminação é realizado como segue: (i) faz-se a clusterização via k-means com todas as variáveis disponíveis, (ii) computa-se a qualidade da clusterização através do SI médio, (iii) elimina-se a variável com menor índice de importância no banco de variáveis remanescentes, retornando ao passo (i). Essas iterações serão realizadas até restar apenas uma variável.

2.3.4 Passo 4 – Construir um perfil relacionando SI médio versus número de variáveis

Na sequência, gera-se um gráfico relacionando o valor de SI médio com o percentual de variáveis retidas, conforme ilustrado na Figura 2.1. O subconjunto de variáveis a ser retido é aquele responsável pelo máximo SI médio. No caso de dois ou mais subconjuntos conduzirem ao mesmo SI, deve-se optar por aquele que retém menos número de variáveis.

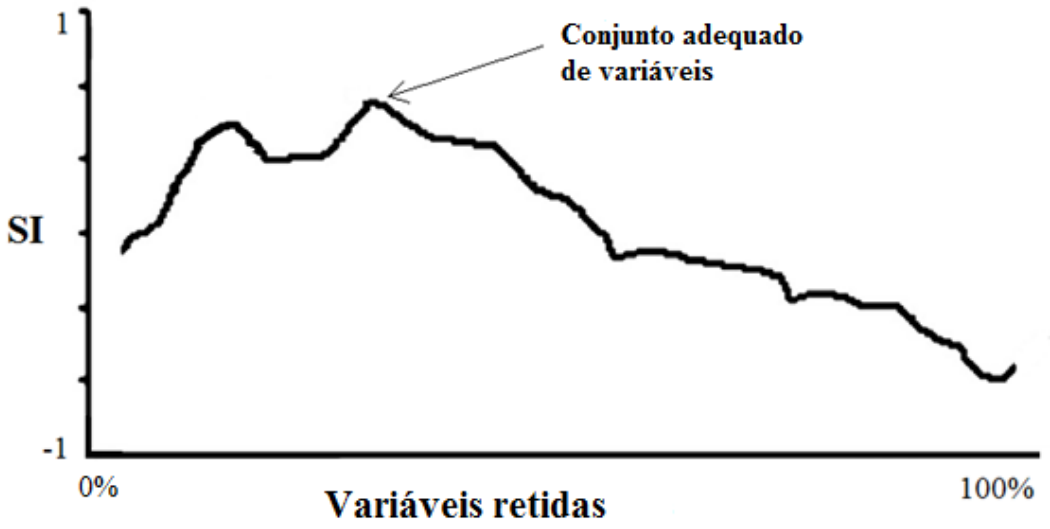

Figura 2.1: Perfil do SI médio com a eliminação das variáveis

2.4 Resultados e Discussão

Foram coletados dados gerados da análise por ATR-FTIR de 69 amostras de Viagra® e 120 amostras de Cialis®. Do total das amostras de cada medicamento tem-se 21 amostras autênticas comerciais de Viagra® e 12 amostras de Cialis®, adquiridas através dos laboratórios Pfizer Ltda. e Eli Lilly do Brasil Ltda. e em farmácias locais de Porto Alegre – RS; e as demais, amostras falsificadas, fornecidas pelo departamento da Polícia Federal de Porto Alegre – RS. A análise de perfil por espectroscopia de infravermelho ATR-FTIR gerada para o Viagra® é apresentada na Figura 2.2; onde cada linha representa uma amostra do medicamento que foi analisada na região espectral de infravermelho médio que vai de 525 a 1800 cm^{-1}, pois, de acordo com Ortiz *et al.* (2013) nesta região contém as características de absorção dos principais componentes, permitindo uma melhor detecção de diferenças nos espectros; para cada amostra analisada obteve-se 661 variáveis. Os passos (2) a (4) do método proposto foram operacionalizados através do *software MATLAB*®, versão R2012b. É importante enfatizar que dois clusters são formados na sistemática proposta para cada banco de dados, objetivando agrupar as amostras de acordo com sua procedência (autêntica ou falsa).

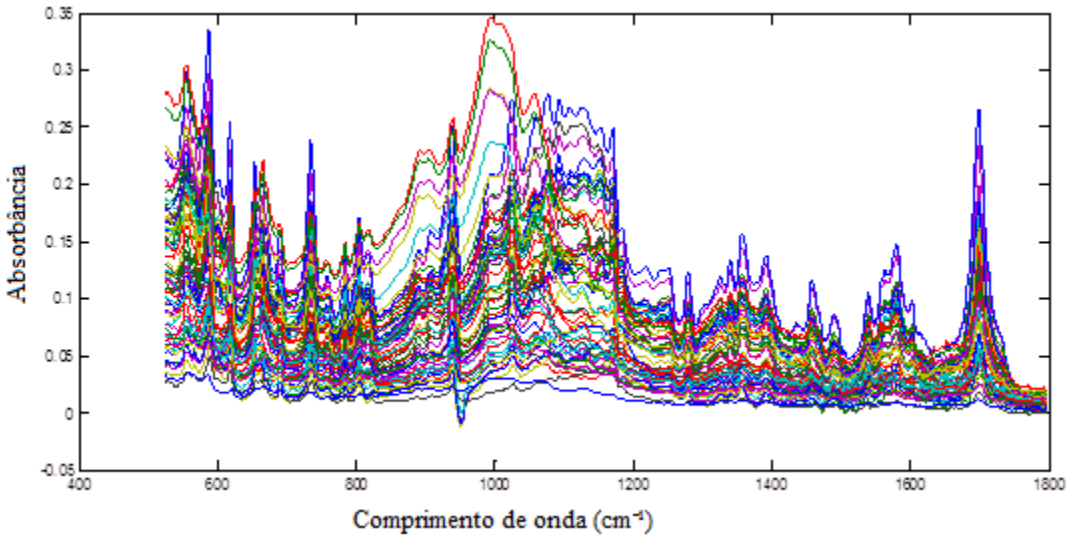

Figura 2.2: Gráfico da análise ATR-FTIR absorbância *vs.* Comprimento de Onda (variáveis) do Viagra®

Para cada banco de dados foram obtidos três resultados decorrentes da aplicação dos três índices de importância testados. O processamento dos dados do Viagra® para a rotina proposta no método teve um tempo inferior a 16 segundos em cada índice testado, já o tempo

de processamento do Cialis® foi de aproximadamente 26 segundos para cada índice. Tais valores são justificados pelo maior número de obervações do Cialis® frente ao Viagra®.

As Figuras 2.3 e 2.4 apresentam o gráfico que relaciona o SI médio com o percentual de variáveis retidas quando o índice v_{j2} é aplicado nos dados do Viagra® e do Cialis®, respectivamente. Através dessas figuras é visualizado o aumento do SI médio à medida que as variáveis, tidas como menos importantes pelo índice são retiradas da análise. O subconjunto de variáveis recomendado é aquele responsável pelo máximo SI médio, observado no ponto indicado pela seta. Tal valor corresponde a 0,8242 para o Viagra®, 0,8720 para o Cialis® e apenas uma variável retida em cada análise.

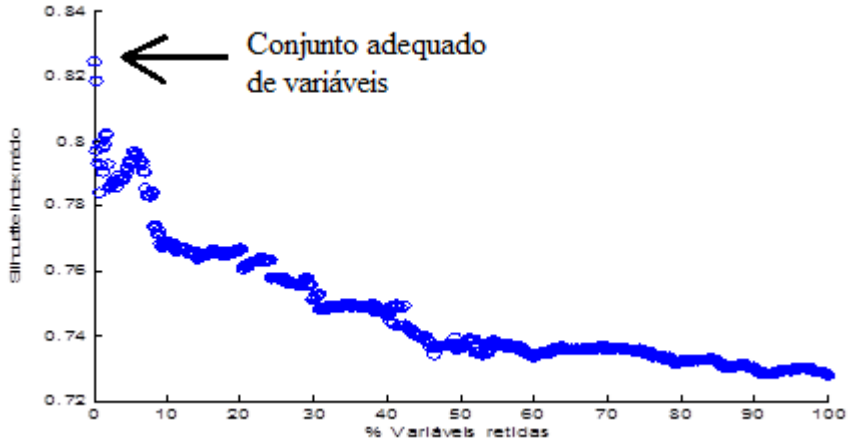

Figura 2.3: Gráfico do SI médio *vs.* porcentagem de variáveis retidas do índice v_{j2} aplicado ao Viagra®

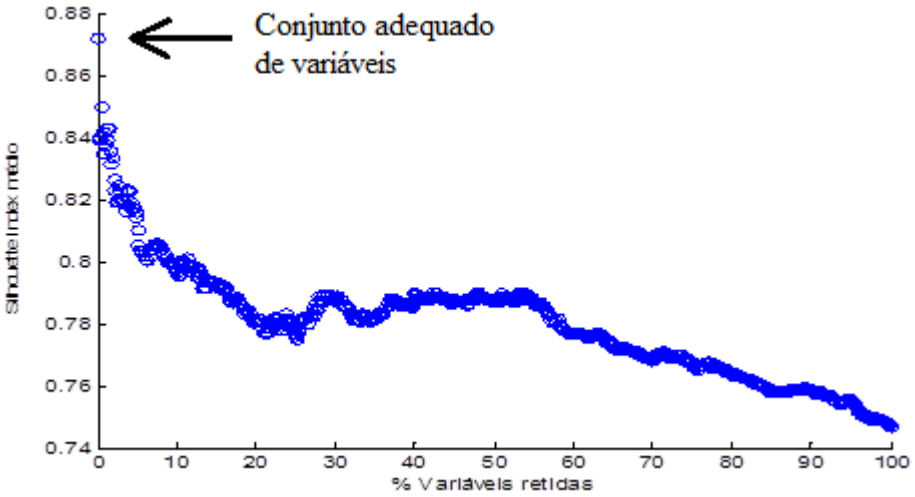

Figura 2.4: Gráfico do SI médio *vs.* porcentagem de variáveis retidas do índice v_{j2} aplicado ao Cialis®

Na Tabela 2.1 são apresentados os valores de SI médio para dois *clusters* (alusivos as amostras autênticas e falsificadas) e o percentual de variáveis retidas para cada índice de importância de variável proposto. Percebe-se que os índices v_{j2} e $v_{j3,}$ que incluem o parâmetro de variância explicada (λ_a) em suas formulações, conduziram a valores de SI médio superiores ao índice v_{j1} e a um menor percentual de variáveis retidas. Nota-se também que os índices v_{j2} e v_{j3} apresentam resultados diferentes, mostrando que há diferença relevante de resultado quando os pesos gerados pela ACP são multiplicados pela variância explicada ou elevados a essa variância.

Para os dados do Viagra®, nota-se que o índice v_{j1}, retém 10,28% das variáveis originais; em contrapartida, os índices v_{j2} e v_{j3}, retêm 0,15% e 2,26%, respectivamente, das variáveis originais. Já para o Cialis® todos os índices retêm menos de 1% das variáveis originais, porém o SI médio melhora em aproximadamente 3% ao retirar duas variáveis da análise, como demonstrado ao comparar os valores de v_{j1} e v_{j2}.

Tabela 2.1 – Valores de SI médio e número de variáveis retidas em cada índice de importância de variável

		Índice			Sem índice
		vj (1)	**vj (2)**	**vj (3)**	
VIAGRA	**SI médio**	0,7928	0,8242	0.8054	0,7277
	nº variáveis retidas	68	1	15	-
	% variáveis retidas	10,2874	0,1513	2,2693	-
CIALIS	**SI médio**	0,8473	0,872	0,8778	0,7469
	nº variáveis retidas	3	1	1	-
	% variáveis retidas	0,4539	0,1513	0,1513	-

A representação gráfica do SI de cada observação alocada ao seu respectivo *cluster*, utilizando o índice v_{j2} aplicado ao Cialis®, é apresentada na Figura 2.5. Nota-se que todas as observações apresentam valores positivos de SI, mostrando que nenhuma amostra foi alocada erroneamente no seu *cluster*. Os gráficos SI vs. *Cluster* gerados pelos índices v_{j2} e v_{j3} para o Viagra® e Cialis® foram semelhantes ao apresentado na Figura 2.5.

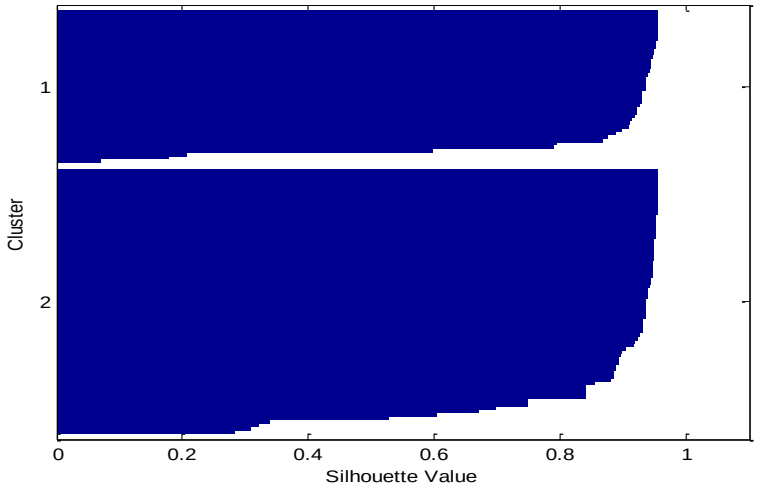

Figura 2.5: Gráfico do SI *vs.* cluster após a utilização do índice de importância de variável v_{j2} para o Cialis®

Na Figura 2.6 é apresentada a clusterização do banco de dados do Cialis® utilizando todas as variáveis. Pelo valor negativo de SI apresentado no *cluster* 2, é visível a existência de amostras alocadas erroneamente ao *cluster*; observa-se ainda que os valores de SI são menores em cada *cluster* quando comparado aos valores da Figura 2.5. Isto demostra que a seleção de variáveis contribui na clusterização mais consistente das amostras, visto que reduziu alocações erradas das amostras e aumentou a pertinência das amostras alocadas aos *clusters* finais.

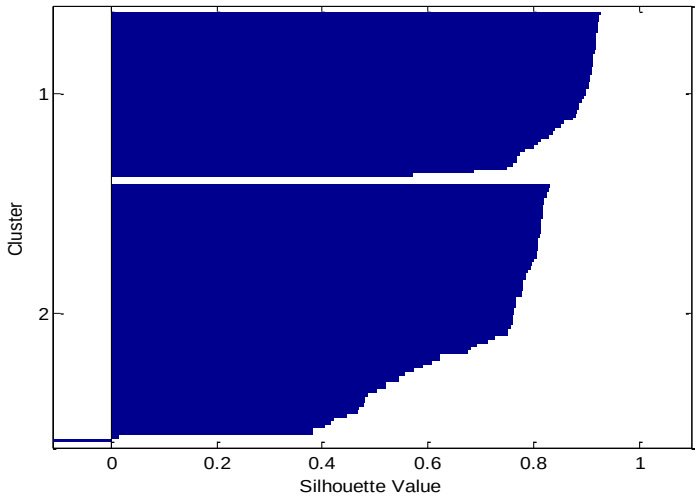

Figura 2.6: Gráfico do SI *vs.* cluster utilizando todas as variáveis do banco de dados do Cialis®

2.5 Conclusões

Este artigo propôs um método de seleção de variáveis para a clusterização de amostras de medicamentos. Tal abordagem é justificada pelo fato que a inserção de todas as variáveis na formação de agrupamentos pode gerar uma perda de qualidade do *cluster* devido à existência de variáveis ruidosas e pouco relevantes. O método é apoiado na análise de componentes principais para a criação de três índices de importância de variáveis. Cada índice, quando aplicado ao banco de dados, orienta um processo iterativo de eliminação de variáveis do tipo *backward*, onde a variável com menor índice de importância é retirada e uma nova clusterização é executada até restar apenas uma variável. Em cada clusterização é obtido o valor do SI médio, que calcula a qualidade do agrupamento, e constrói-se um gráfico relacionando o SI médio com o número de variáveis retidas para identificar o conjunto de variáveis adequado.

O método foi aplicado na formação de dois *clusters* (alusivos a classes autênticas e falsificadas) em bancos de dados do Viagra® e Cialis®. Observou-se que os índices de importância de variáveis, v_{j2} e v_{j3}, que se apoiam no parâmetro variância explicada (λ_a), geram resultados mais consistentes de clusterização, com maiores valores de SI médio e uma menor porcentagem de variáveis retidas.

Sugere-se, para trabalhos futuros, a aplicação do método proposto neste artigo apenas nas amostras falsificadas com vistas à obtenção de *clusters* que indiquem similaridades entre as fontes falsificadoras dos medicamentos. Outro tema possível é propor e testar novos índices de importância de variáveis apoiados em outras técnicas multivariadas.

2.6 Referências Bibliográficas

ANDERSON, T. W. *Na introduction to multivariate statistical analysis*. 3ª ed. New Jersey: John Wiley & Sons, Inc. Hoboken, 2003.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. *Selecting the best variables for classifying production batches into two quality levels*. Chemometrics and Intelligent Laboratories Systems, v. 97, p. 111-117, 2009.

ANZANELLO, M. J.; FOGLIATTO, F. S. *Selecting the best clustering variables for grouping mass-customized products involving workers' learning*. International Journal of Production Economics, v. 130, p. 268-276, 2011.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGER, 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, p. 209-214, 2013.

BARROS, A. S.; RUTLEDGE, D. N. *PLS_Cluster: a novel technique for cluster analysis*. Chemometrics and Intelligent Laboratories Systems, v. 70, p. 99-112, 2004.

BRUSCO, M. J. *Clustering binary data in the presence of masking variables*. Psychological Methods, v. 9, p. 510-523, 2004.

BRUSCO, M. J.; CRADIT, J. D. *A variable-selection heuristic for k-means clustering*. Psychometrika, v. 66, p. 249-270, 2001.

CARLSON, R.; GAUTUN, H. *Combinatorial libraries and the development of organic synthetic methods. PLS modeling to discriminate between successful and failed reaction systems*. Chemometrics and Intelligent Laboratories Systems, v. 78, p. 113-124, 2005.

CERVO, V. L.; ANZANELLO, M. J. *Avaliação da robustez de uma sistemática de seleção de variáveis para clusterização através de experimentos de simulação*. São Paulo: Revista Gestão e Produção, 2013.

DE SOETE, G. *Optimal variable weighting for ultrametric and additive tree clustering*. Quality and Quantity, v. 20, p. 169-180, 1986.

DING, C.; HE, X. *K-means clustering via principal component analysis*. In: Russ Greiner, Dale Schuurmans (Eds.). Proceedings of the 21st International Machine Leraning Conference, ACM Press, p. 225-232, 2004.

DOWNS, G. M.; BARNARD, J. M. *Clustering methods and their uses in computational chemistry*. Reviews in Computational Chemistry, vol. 18,p. 1–40, 2002.

DUDA, R.; HART, P.; STORK, D. *Pattern Recognition*. 2ª ed. New York: Willey, 2001.

FERNANDEZ, F. M.; HOSTETLER, D.; POWELL, K.; KAUR, H.; GREEN, M.; MILDENHALL, D. C.; NEWTON, P. N. *Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries*. Analyst, v. 136, p. 3073-3082, 2011.

GNANADESIKAN, R.; KETTENRING, J.; TSAO, S. *Weighting and selection of variables for cluster analysis*. Journal of Classification, v. 12, p. 113-136, 1995.

GAUDIANO, M.C.; DI MAGGIO, A.; ANTONIELLA, E.; VALVO, L.; BERTOCCHI, P.; MANNA, L.; BARTOLOMEI, M.; ALIMONTI, S.; RODOMONTE, A. L. *An LC method for the simultaneous screening of some common counterfeit and sub-standard antibiotics Validation and uncertainty estimation.* Journal of Pharmaceutical and Biomedical Analysis, v. 48, p. 303-309, 2008.

HAIR JR., J. F. *Análise multivariada de dados*. 5ª ed. Porto Alegre: Bookman, 2005.

HIND, A. R.; BHARGAVA, S. K.; MCKINNON, A. *At the solid/liquid interface: FTIR/ATR – the tool of choice*. Advances in Colloid and Interface Scienc, v.93, p. 91-114, 2001.

HOLZGRABE, U.; MALET-MARTINO, M. *Analytical challenges in drug counterfeit-ing and falsification—the NMR approach*. Journal of Pharmaceutical and Biomedical Analysis, v. 55, p. 679-687, 2011.

HONDA, K.; NOTSU, A.; ICHIHASHI, K. *PCA-guided k-means with variable weighting and its application to document clustering*. Lectures Notes in Computer Science, v. 5861, p. 282-292, 2009.

HUANG, J. Z.; NG, M. K.; RONG, H.; LI, Z. *Automated variable weighting in k-means type clustering*. IEEE Transaction on Pattern Analysis and Machine Intelligence, v. 27, p. 657-668, 2005.

JOLLIFFE, I. T. *Principal Component Analysis*. 2ª ed. New York: Springer, 2002.

JUNG, C. R.; ORTIZ, R. S.; LIMBERGER, R.; MAYORGA, P. *A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis.* Forensic Science International, v. 216, p. 92-96, 2012.

KAUFMAN, L.; ROUSSEEUW, P.; *Finding groups in data: an introduction to cluster analysis*. New Jersey: Wiley Interscience, 2005.

KIM, H. C.; KIM, D.; BANG, S. Y. *An efficient model order selection for PCA mixture model*. Pattern Recognition Letters, v. 24, p. 1385-1393, 2013.

LATIFOĞLU, F.; POLAT, K.; KARA, S.; GÜNES, S. *Medical diagnosis os atherosclerosis from Carotid Artery Doppler Signals using principal componente analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS)*. Journal of Biomedical Informatics, v. 41, p. 15-23, 2008.

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, p. 189-200, 2002.

LOPES, M. B.; WOLFF, J-C. *Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging*. Analytica Chimica Acta, v. 633, p. 149-155, 2009.

MAKARENKOV, V.; LEGENDRE, P. *Optimal variable weighting for ultrametric and additive trees and k-means partitioning: methods and software*. Journal of Classification, v. 18, v. 245-271, 2001.

MEDEIROS, A. R. B. *Uso de ATR/FTIR e FTNIR associado a técnicas quimiométricas para a quantificação de aditivos em gasolina automotiva*. Tese de Dissertação, Brasília, 2009.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SAEBO, S. *A review of variable selection methods in Partial Least Squares Regression*. Chemometrics and Intelligent Laboratory Systems, v. 118, p. 62-69, 2012.

MODHA, D.S.; SPANGLER, W.S. *Feature weighting in k-means clustering*. Machine Learning, v. 47, 2002.

ORTIZ, R. S.; MARIOTTI, K. C.; FANK, B.; LIMBERGER, R. P.; ANZANELLO, M. J.; MAYORGA, P. *Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the same pharmaceutical powder mixture be used to falsify two medicines?* Forensic Science International, v. 226, p. 282-289, 2013.

PLAEHN, D.; LUNDAHL, D. S. *An L-PLS preference cluster analysis on French consumer hedonics to fresh tomatoes*. Food Quality and Preference, v. 17, p. 243-256, 2006.

POON, L. K. K.; ZHANG, N. L.; LIU, A. H. *Model-based clustering of high-dimensional data: Variable selection versus facet determination*. International Journal of Approximate Reasoning, v. 54, p. 196-215, 2013.

RENCHER, A. C. *Methods of multivariate analysis*. 2ª ed. New York: Wiley-Interscience, 2002.

SACRÉ, P.Y.; DECONINCK, E.; DE BEER, T.; COURSELLE, P.; VANCAUWEBERGHE, R.; CHIAP, P.; CROMMEN, J.; DE BEER, J. O. *Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines*. Journal of Pharmaceutical and Biomedical Analysis, v. 53, p. 445-453, 2010.

SEBZALLI, Y. M.; WANG, X. Z. *Knowledge discovery from process operational data using PCA and fuzzy clustering*. Artificial Intelligence, v. 14, p. 607-616, 2001.

STEINLEY, D.; BRUSCO, M. J. *A new variable weighting and selection procedure for k-means cluster analysis*. Multivariate Behavioral Researche, v. 43, p. 77-108, 2008.

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

VICINI, L.; SOUZA, A. M. *Análise Multivariada da teoria a prática*. Santa Maria: UFSM, CCNE, 2005.

WHO Media centre, *Medicines: spurious/falsely-labelled/ falsified/counterfeit (SFFC) medicines.* (Fact sheet nº 275), 2012. Disponível em < http://www.who.int/mediacentre/factsheets/fs275/en/index.html>. Acesso em 30 de agosto de 2013.

XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J. W.; HOLMES, M.; HANPIN, M. *Variables selection methods in near-infrared spectroscopy.* Analytica Chimica Acta, v. 667, p. 14-32, 2010.

XUE, J.; LEE, C.; WAKEHAM, S. G.; ARMSTRONG, R. A. *Using principal components analysis (PCA) with cluster analisys to study the organic geochemistry of sinking particles in the ocean*. Organic Geochemistry, v. 42, p. 356-367, 2011.

XU, L.; JIANG, J.-H.; WU, H.-L.; SHEN, G.-L.; YU, R.-Q. *Variable-weighted PLS*. Chemometrics and Intelligent Laboratory Systems, v. 85, p. 140-143, 2007.

YÜCEL, Y.; SULTANOĞLU, P. *Determination of industrial pollution effects on citrus honeys with chemometric approach*. Food Chemistry, v. 135, p. 170-178, 2012.

3 Segundo Artigo: Seleção de variáveis com vistas à classificação de medicamentos apoiada em Algoritmo Genético

Gabrielli Harumi Yamashita

Michel José Anzanello

Resumo

Para identificar medicamentos fraudulentos, cada vez mais tem-se utilizado a análise de perfil por espectroscopia de infravermelho (ATR-FTIR). No entanto, essa técnica analítica tipicamente gera um grande número de variáveis, tornando importante a utilização de técnicas que selecionem as variáveis mais relevantes. Este artigo apresenta um método de seleção de variáveis para a classificação de amostras de medicamentos em duas classes (originais ou falsificados). O Algoritmo Genético (AG) gera subconjuntos de variáveis utilizadas na classificação das amostras via *k*NN (*k-nearest neighbor*). A classificação é repetida diversas vezes para cada subconjunto, de modo a obter uma acurácia média e definir o subconjunto a ser retido com base na acurácia de classificação. O método foi aplicado em bancos de ATR-FTIR de Cialis® e Viagra®; observou-se que a acurácia média aumenta em até 8% para o Cialis® e até 20% para o Viagra® quando se realiza a classificação com menos de 2% das variáveis originais.

Palavras-chave: ATR-FTIR, Seleção de variáveis, *k*NN, Algoritmo Genético.

3.1 Introdução

A falsificação de medicamentos é uma ação criminosa e prejudicial à saúde do consumidor, por isso percebe-se o aumento de estudos voltados à identificação desses medicamentos (JUNG *et al*., 2012; LOPES e WOLFF, 2009). Devido ao aumento de práticas de falsificação e com o intuito de combatê-las a *World Health Organization* (WHO) criou, em 2006, um grupo chamado IMPACT (*International Medical Products Anti-Counterfeiting Taskforce*), que busca soluções para diminuir e até eliminar a produção desses medicamentos fraudulentos, além de sensibilizar a população mundial sobre os riscos da utilização desses produtos (WHO, 2008).

Os estudos sugerem que o aumento da falsificação de medicamentos pode estar associado à fiscalização deficiente, legislação inadequada, fácil acesso às tecnologias necessárias aos falsificadores para copiar os produtos originais, alto custo dos medicamentos originais e ignorância dos consumidores em relação aos riscos de saúde que medicamentos adulterados podem trazer. Tais produtos podem ter sido manipulados em ambientes insalubres, conter concentrações inapropriadas dos ingredientes e serem mal rotulados (WHO, 2008; FERNANDEZ *et al*., 2011, LEBEL *et al*., 2014).

Além de consequências para a saúde dos usuários, o impacto de medicamentos falsificados na economia global é substancial (DEGÁRDIN *et al*., 2011). Deconink (2012) cita que a internet é o meio mais usado para a comercialização desses medicamentos na Bélgica, onde 50% dos medicamentos adquiridos são falsificados. No Brasil, segundo os dados da Policia Federal (PF), a apreensão de medicamentos fraudulentos tem crescido substancialmente. Um levantamento no banco de dados da PF no período de janeiro de 2007 a setembro de 2010 concluiu que, dos laudos emitidos pelos peritos criminais, 80% dos medicamentos apreendidos eram falsificações do Cialis® e Viagra® (AMES e SOUZA, 2012; JUNG *et al*,. 2012). Esses medicamentos são receitados para combater a disfunção erétil masculina e seu alto índice de falsificação pode ser explicado pelo custo elevado do produto original, uma alta demanda da população brasileira e a facilidade de comprar pela internet (AMES e SOUZA, 2012).

Por conta de tais prejuízos, percebe-se um grande interesse em métodos analíticos capazes de detectar as falsificações de forma rápida e eficiente (DEGÁRDIN *et al*., 2011). Dentre tais métodos, destaca-se o perfil por espectroscopia de infravermelho ATR-FTIR (*Attenuated Total Reflectance - Fourier Transform Infrared*), entendido como um método rápido e confiável (ORTIZ *et al*., 2013). No entanto, tal análise é caracterizada por gerar resultados com um número elevado de variáveis em relação ao número de amostras, tornando necessária a seleção das variáveis mais relevantes e menos ruidosas, que podem causar distorções aos modelos gerados (ANZANELLO *et al*., 2013; XIAOBO *et al*., 2010).

O principal objetivo da seleção de variáveis é identificar um subconjunto reduzido das variáveis originais responsável pelas informações mais importantes a serem incluídas em um modelo preditivo ou de classificação (RAYMER *et al*., 2000). A seleção de variáveis apoiada no Algoritmo Genético (AG) tem sido vastamente utilizada devido à sua eficiência, robustez e versatilidade (GOLDBERG, 1998). O AG é um algoritmo que se baseia na seleção natural

para chegar a um conjunto de variáveis que produza um resultado consistente em termos de predição ou classificação (WIEGAND *et al*., 2009).

Este artigo testa um método de seleção das variáveis espectroscópicas mais relevantes utilizando o Algoritmo Genético. Nas proposições deste artigo, são gerados, por meio do AG, subconjuntos de variáveis do banco de dados original e então classificadas as amostras (em originais ou falsificadas) através da ferramenta *k*NN (*k-nearest neighbor*). Cada subconjunto passa pelo processo de classificação várias vezes seguidas de forma a obter uma acurácia média, sinalizando a precisão da classificação com as variáveis retidas. À medida que o AG seleciona os subconjuntos de variáveis, as amostras são classificadas e obtido o valor da acurácia média referente a cada subconjunto. Esse processo se repete de forma a obter o conjunto de variáveis que resultem num maior valor de acurácia média.

Este artigo está estruturado como segue, além desta introdução. Na seção 2 é apresentado um referencial teórico sobre Algoritmo Genético e classificação via *k*NN, na seção 3 é descrito o método proposto. Na seção 4 são apresentados os resultados obtidos em conjunto com as discussões referentes. A seção 5 traz as considerações finais.

3.2 Referencial Teórico

3.2.1 Algoritmo Genético

Inspirado na teoria da evolução das espécies de Darwin, com o objetivo de explicar o processo de adaptação da seleção natural e desenvolver sistemas artificiais que reproduzam computacionalmente os mecanismos da seleção natural, o pesquisador John Holland propôs um algoritmo matemático para a otimização de sistemas complexos, chamado de Algoritmo Genético (AG) (GOLDBERG, 1998). O princípio fundamental do AG é que as gerações derivadas serão mais evoluídas do que seus antecedentes; de tal forma, indivíduos melhores continuariam existindo enquanto indivíduos mais frágeis tenderiam a ser eliminadas (KONZEN *et al.*, 2003).

O algoritmo genético básico funciona de modo iterativo sobre um conjunto de possíveis soluções para o problema em questão, sendo estruturado em seis etapas: geração da população inicial, avaliação da população, teste de convergência ou critério de término,

seleção e aplicação dos operadores do AG (representa duas etapas), e criação de uma nova população (COSTA FILHO E POPPI, 1999). Tais etapas são detalhadas na sequência.

O AG é iniciado com a geração de uma população inicial, onde cada indivíduo é composto por um grupo de genes (variáveis) codificados, usualmente pelo código binário (KIM *et al*., 2007). Cada indivíduo da população corresponde a um ponto no espaço de busca e uma possível solução para o problema em questão (KIM *et al*., 2007). A população é constituída por um número finito de indivíduos e cada um é avaliado através da função de aptidão (função objetivo do problema), que calcula a aptidão de determinado indivíduo em relação ao modelo e avalia se o indivíduo está apto a reproduzir e permanecer na população (GOLDBERG, 1998; KONZEN *et al.*, 2003).

A terceira etapa do algoritmo consiste em determinar um critério de parada, o qual pode ser o tempo de processamento, resultado ótimo alcançado, número de geração do algoritmo ou algum outro indicador definido pelo analista (KELLY e DAVIS, 1991). É necessário definir o critério de parada para garantir a rapidez e a finalização do algoritmo, pois enquanto o AG não satisfizer o critério de parada ele persistirá na busca do resultado ótimo (COSTA FILHO E POPPI, 1999).

A sequência do algoritmo ocorre através da aplicação dos operadores genéticos de seleção, cruzamento e mutação, que guiará a criação de novos indivíduos e potencialmente melhores soluções (LAVINE *et al*., 2002). Na fase da seleção os indivíduos serão escolhidos, de acordo com o valor de aptidão, para posterior cruzamento. O principal objetivo é enfatizar os melhores cromossomos, tentando garantir que os bons indivíduos se reproduzam e gerem indivíduos melhores (ENGELBRECHT, 2008). Entre as técnicas de seleção mais utilizadas estão: (i) roleta, onde cada indivíduo tem a probabilidade de ser selecionado proporcional ao seu valor de aptidão, assim os indivíduos que possuem uma alta aptidão ocuparão uma porção maior na roleta do que os indivíduos que possuem uma aptidão menor, e então a roleta é girada e é selecionado um individuo que participara do processo de geração da nova população, esta ação se repete até selecionar a quantidade desejada de indivíduos; (ii) torneio, que entre dois indivíduos aleatórios seleciona o que tiver o melhor valor em um critério a ser determinado, realiza-se quantas disputas for necessária para selecionar a quantidade de indivíduos que irá reproduzir; e (iii) elitismo, que preserva um grupo de indivíduos mais aptos da população atual para completar a nova população (GOLDBERG, 1998).

Os operadores cruzamento e mutação modificam a população ao longo das gerações, pois são responsáveis por fazer as mudanças de genes dos cromossomos, sendo essencial para a diversificação da população e permitindo a preservação das características relevantes adquiridas dos antecessores (KONZEN *et al.*, 2003; IZQUIERDO, 2013).

O cruzamento é a etapa de reprodução dos indivíduos, onde ocorre a troca de genes entre dois indivíduos da população, gerando novos indivíduos com tendência de herdarem as características dominantes dos seus geradores e transmiti-las para gerações futuras (ENGELBRECHT, 2008). Segundo Costa Filho e Poppi (1999) o cruzamento pode acontecer de três formas: (i) heterossexual, onde faz-se a distinção do gênero de um indivíduo e apenas indivíduos de gêneros diferentes podem cruzar entre si; (ii) homossexual, onde ocorre cruzamento entre indivíduos sem distinção de gênero; e (iii) assexuado, onde ocorre a troca de genes dentro do próprio indivíduo. O cruzamento é responsável pela convergência ao resultado ótimo do modelo, pois se observa que, após algumas gerações, há uma alta taxa de indivíduos que possuem a presença de genes dominantes (COSTA FILHO e POPPI, 1999).

A mutação é uma alteração aleatória no gene de um indivíduo, e ocorre, com uma probabilidade mínima, com o intuito de prevenir o modelo de uma eventual perda de informação relevante que possa ocorrer na fase de cruzamento (GOLDBERG, 1998). Após a aplicação dos operadores tem-se uma nova população e as etapas são refeitas até satisfazer o critério de parada (COSTA FILHO E POPPI, 1999).

As vantagens apresentadas pelo algoritmo genético frente a outras técnicas de otimização, segundo Goldberg (1998) e Izquierdo (2013), são: fácil implementação; proporciona maior flexibilidade no tratamento do problema a ser resolvido; é mais resistente a se prender em ótimos locais; utiliza regras de transição probabilísticas e não determinísticas; não requer informações dos gradientes da superfície definida pela função objetivo; é robusto e aplicável a uma grande variedade de problemas; e é facilmente combinável com outras técnicas e heurísticas.

3.2.2 Seleção de variáveis utilizando o Algoritmo Genético

A seleção de variáveis é importante por diversas razões, dentre as quais melhorar o desempenho preditivo e classificatório do modelo proposto, gerar modelos mais robustos e

mais confiáveis, e tornar um modelo mais fácil para a compreensão e manuseio pelos usuários (LEARDI *et al*., 2002). Selecionar variáveis pode ser considerado um problema de otimização, e nesses procedimentos a aplicação do algoritmo genético tem sido bastante eficiente por sua capacidade de buscar vários pontos em paralelo, podendo encontrar um ótimo global e evitando cair em um ótimo local (RAYMER *et al*., 2000; LIU e ONG, 2008).

Quando se trabalha a seleção de variáveis espectroscópicas no AG, a codificação se dá de forma que cada gene do cromossomo representa uma das variáveis do espectro, fazendo com que o cromossomo contenha todas as variáveis do espectro (COSTA FILHO E POPPI, 1999). O código binário é utilizado, e cada gene pode receber o valor 1 ou 0, onde 1 representa que a variável está selecionada e 0 representa que a variável não está inserida no modelo (RAYMER *et al*., 2000).

O AG, segundo Basgalupp (2007), pode ser empregado de duas formas na seleção de variáveis: *wrapper* e *filter*. A primeira ocorre quando é necessário introduzir um classificador para o cálculo da função de aptidão, tornando o processo mais demorado, pois cada candidato a solução precisa passar por um classificador; caso não seja necessária a indução do classificador o método pertence ao tipo *filter*.

Na maior parte dos estudos reportados pela literatura, os AGs são utilizados com a função *wrapper*, como visto em Leardi *et al*. (2002), que combinam o algoritmo genético com a regressão por mínimos quadrados parciais (PLS) para a previsão da concentração de aditivo em filmes de polímeros. O AG gera conjuntos de variáveis que são utilizados na predição PLS e tem como resultado a estimativa do erro preditivo; o subconjunto de variáveis responsável pelo menor erro de predição é tido como a melhor solução. As variáveis selecionadas naquele estudo são consistentes em termos químicos quando validadas por especialistas, demostrando que o AG é eficaz em selecionar variáveis corretas de forma automatizada. Já Raymer *et al*. (2000) e Kelly Jr e Davis (1991) utilizam o AG para otimizar um vetor de pesos para as variáveis e assim estabelecer os conjuntos prováveis de variáveis relevantes para a classificação; em seguida, o banco de dados é classificado pela ferramenta *k nearest neighbor* (*k*NN) e é selecionado o subconjunto de variáveis que geraram a máxima acurácia. Os autores concluem que, ao utilizar a seleção de variáveis, a classificação apresenta um incremento no desempenho de classificação.

Liu e Ong (2008) reforçam que a seleção de variáveis através do AG pode efetivamente encontrar a melhor solução global e obter precisão no modelo proposto. Os autores objetivam determinar o número ideal de *clusters* na segmentação do mercado; para tanto, utilizam o AG na seleção de variáveis em conjunto com a técnica de clusterização *k-means*. Por sua vez, Will *et al*. (2013) utilizam o AG para resolver o problema de seleção de variáveis para a estimativa de radiação solar, apresentando uma metodologia eficaz para estimar uma variável quando o banco de dados tem falta de informações ou selecionar variáveis de entrada relevantes mudando apenas os parâmetros do AG. Por fim, Aladag *et al*. (2014) mostram que a seleção de variáveis através do AG pode ser usada também na construção de modelos *fuzzy* de previsão de séries temporais para descartar as variáveis defasadas do modelo. Não foram identificadas, ao final da revisão, estudos utilizando AG na seleção de variáveis espectroscópicas de Cialis® e Viagra®.

3.2.3 *k Nearest Neighbor* (*k*NN)

O *k*NN é uma das ferramentas mais utilizadas para a classificação de observações. Ele categoriza as amostras em classes apoiando-se na Distância Euclidiana de uma nova amostra em relação aos seus *k* vizinhos mais próximos. Uma nova amostra é alocada a uma classe já conhecida se a maioria dos seus *k* vizinhos pertencer àquela classe (WAN *et al*., 2012; ANZANELLO *et al*., 2013). O valor de *k* é obtido por testes que maximizem a precisão, confiabilidade ou sensibilidade da classificação no conjunto de treinamento onde a classe de cada amostra é conhecida (ANZANELLO *et al*., 2009).

Esta técnica utiliza a fase de treinamento para mapear as amostras conhecidas em regiões com maior semelhança entre suas variáveis; quando uma nova amostra dá entrada no classificador, este utiliza a distância entre essa nova amostra e as classes que foram determinadas na fase de treinamento para alocar o novo ponto à sua categoria específica (HAN *et al*., 2001). Em suma, o *k*NN faz uma decisão comparando uma nova amostra sem classe específica com os dados de base que já estão em classes determinadas (CHAOVALITWONGSE *et al*., 2007).

O *k*NN tem sido amplamente utilizado em estudos por ser conceitualmente simples, precisar de um pequeno número de dados de amostras para treinamento, requer apenas um

parâmetro *k* (que dentro de uma faixa razoável tipicamente não interfere significativamente na precisão da classificação), e estar disponível em vários pacotes de *software* (HAN *et al*., 2001; ANZANELLO *et al*., 2011). Anzanello *et al.* (2013) utilizam o *k*NN para classificar amostras de medicamentos para disfunção erétil em autênticos ou falsificados. Após a classificação, são obtidas a acurácia, sensibilidade e especificidade, as quais apontam que o método de seleção de variáveis proposto aumenta a precisão da classificação em aproximadamente 2%.

Outras aplicações práticas reforçam o bom desempenho do *k*NN, como em Anzanello *et al*. (2009) na classificação da qualidade de lotes de produção em conformes e não conformes. Li e Zhang (2014) utilizam o método para classificar, ainda na linha de produção, potenciais falhas de semicondutores. Chaovalitwongse *et al*. (2007) aplicam o *k*NN para detectar anomalias na atividade cerebral; Anzanello *et al*. (2011) utilizam para determinar grupos de avaliadores de alimentos que mais se assemelham em suas avaliações; Zuo *et al*. (2014) utilizam para detecção da doença de Parkinson; Weiss *et al*. (1999) na mineração de texto e Lee *et al*. (1991) no reconhecimento de dígitos escritos a mão.

3.3 Método

A necessidade de selecionar as variáveis mais relevantes quando se trabalha com dados espectroscópicos surge devido ao grande número de dados gerados por esta análise; quando inseridos no modelo (seja de predição ou classificação), variáveis irrelevantes podem resultar em distorções e gerar conclusões equivocadas (COSTA FILHO e POPPI, 2002). Com foco em propósitos de classificação de amostras de medicamentos em duas classes (original ou falsificado), o método abordado para a seleção das variáveis é dividido em quatro passos: (1) coletar dados espectroscópicos para análise, (2) definir os parâmetros e critérios do AG, (3) aplicar o AG para gerar subconjuntos de variáveis e utilizar o *k*NN para calcular a função de aptidão, e (4) construir o gráfico relacionando a acurácia média *versus* número de variáveis e identificar a melhor solução. Os passos (3) e (4) serão aplicados para diversos valores de *k* e diferentes proporções de treino e teste do banco de dados, gerando valores médios de desempenho de classificação. Esses passos são detalhados a seguir.

3.3.1 Passo 1 - Coletar dados espectroscópicos para análise

Os dados coletados provêm da análise de perfil por espectroscopia de infravermelho ATR-FTIR de amostras de medicamentos. Segundo Ortiz *et al*. (2013), a técnica ATR-FTIR para análise de amostras de medicamentos é simples, obtém resultados rápidos e apresenta facilidade em manipular as amostras, sem a necessidade de pastilhamento ou um pré-tratamento; de tal forma, torna-se uma ferramenta eficaz para investigações forenses com o objetivo de classificar amostras.

Em um equipamento de espectroscopia no infravermelho por Transformada de Fourier (FTIR), o feixe de radiação é separado em um divisor, onde uma parte vai para um espelho fixo e a outra para um espelho móvel. Após a reflexão, os dois feixes se encontram e sofrem uma interferência após terem percorrido distâncias diferentes devido ao percurso do espelho móvel. A diferença no caminho percorrido pelos dois feixes é chamada de atraso e o gráfico da intensidade da radiação em função do atraso é chamado de interferograma. A transformada de Fourier converte o interferograma que está em função do atraso, para estar em função da frequência, de modo a obter um espectro de frequências (MEDEIROS, 2009). Para fazer a medições FTIR pode-se utilizar a técnica de reflexão total atenuada (ATR), que emprega uma propriedade de reflexão interna total, na qual uma amostra é colocada em contato com um elemento de alto índice de refração; a radiação atravessa o elemento de reflexão sendo refletida e direcionada para um detector (HIND *et al*., 2001).

Os espectros gerados pela análise ATR-FTIR tem como característica serem formados por bandas contínuas em vez de respostas discretas, porém essas faixas contínuas são compostas por muitas respostas discretas em estreita proximidade de comprimento de onda, que são identificadas como as variáveis a serem estudadas (LEARDI *et al*.,2002).

3.3.2 Passo 2 - Definir os parâmetros e critérios do AG

Especificar os parâmetros demandados pelo AG: tamanho da população inicial, probabilidade de cruzamento, probabilidade de mutação e número máximo de variáveis selecionadas. Esses parâmetros serão definidos de acordo com recomendações encontrados na literatura, como em Leardi *et al.* (2002), Raymer *et al*. (2000) e Costa Filho e Poppi (2002). É necessário também definir a função de aptidão que será maximizada e o critério de parada do

algoritmo, que pode ser o número de gerações, o tempo de processamento ou quando atingir um determinado valor de aptidão, que neste estudo é o valor de acurácia.

3.3.3 Passo 3 - Aplicar o AG para gerar subconjuntos de variáveis e utilizar o *k*NN para calcular a função de aptidão

Nesta etapa, dá-se início à execução do AG como método de seleção de variáveis. A Figura 3.1 apresenta o fluxograma desta etapa. A população inicial, definida de acordo com recomendações encontradas na literatura, é gerada randomicamente e cada indivíduo é composto por uma quantidade (definida no passo 2) de variáveis do banco de dados; assim, essas variáveis são selecionadas em todas as amostras e realizada a classificação através da ferramenta *k*NN.

No *k*NN, o banco de dados é dividido em treino e teste, sendo esta proporção definida pelo usuário; as amostras que irão compor os dados de treino e os dados de teste são escolhidas randomicamente pela ferramenta. Para cada subconjunto de variáveis selecionada pelo AG são feitas diversas classificações no *k*NN, de forma a obter resultados de diversas combinações do banco de treino e do banco de teste.

Após as classificações, mede-se a acurácia média, obtida pela média das razões entre o número de classificações corretas e o número total de classificações em cada iteração do *k*NN. Em seguida, verifica-se se o critério de parada foi atendido; se sim, tem-se o melhor subconjunto de variáveis relevantes; caso contrário, é realizada a evolução dos indivíduos através da seleção, cruzamento e mutação, criando uma nova população (formada por variáveis distintas das anteriormente testadas) e repetindo o ciclo. Este procedimento é realizado diversas vezes, alterando os valores de *k* e as proporções de treino e teste do banco de dados (para garantir que a classificação não seja beneficiada por uma separação favorável dos dados em treino e teste).

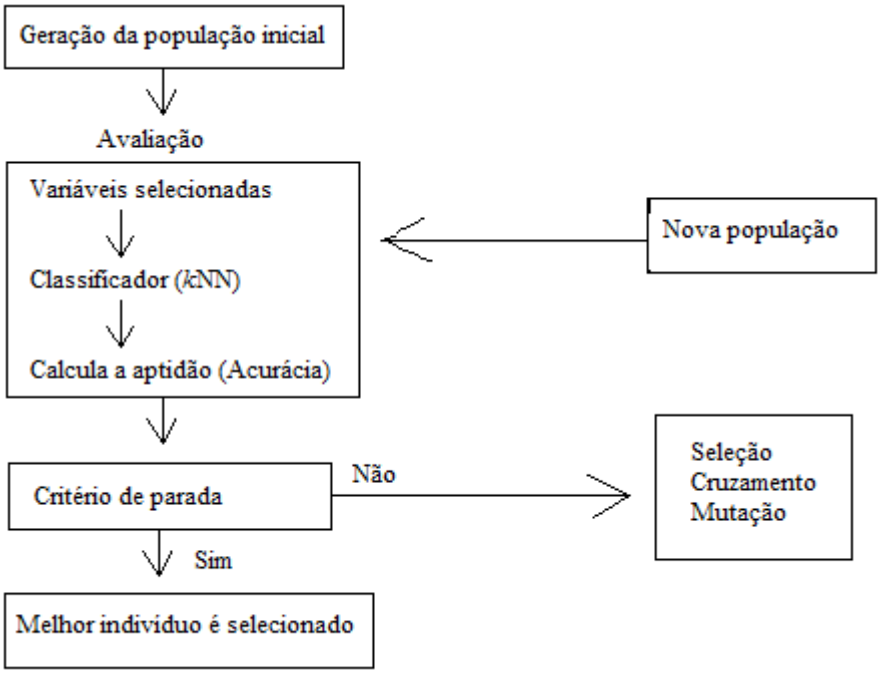

Figura 3.1: Fluxograma do *k*NN integrado ao algoritmo genético

3.3.4 Passo 4 - Construir o gráfico relacionando a acurácia média versus número de variáveis e identificar a melhor solução

Na sequência, é gerado um gráfico associando o valor da acurácia média com a porcentagem de variáveis selecionadas, como mostra a Figura 3.2. Nas configurações do AG é imposto um número máximo de variáveis a ser selecionado, fazendo com que o eixo das abscissas tenha seu valor máximo e dependa desse parâmetro, porém a quantidade de variáveis selecionadas em cada iteração pode variar de 1 até o número máximo. O subconjunto que obtém as variáveis mais representativas é aquele que apresenta o maior valor de acurácia média (função de aptidão).

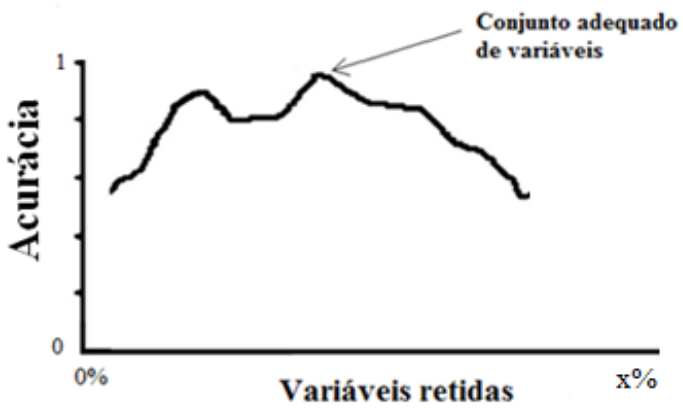

Figura 3.2: Comportamento da acurácia média de acordo com a quantidade de variáveis retidas

3.4 Resultados e Discussões

Foram coletados dados gerados da análise por ATR-FTIR de 69 amostras de Viagra® e 120 amostras de Cialis®. Deste total de amostras de cada medicamento tem-se 21 amostras autênticas comerciais de Viagra® e 12 amostras de Cialis®, adquiridas através dos laboratórios Pfizer Ltda. e Eli Lilly do Brasil Ltda. e em farmácias locais de Porto Alegre – RS; e as demais, amostras falsificadas, fornecidas pelo departamento da Polícia Federal de Porto Alegre – RS. Na Figura 3.3 é apresentado o resultado de uma análise ATR-FTIR em uma amostra comercial do Cialis® e do Viagra®. Cada amostra foi analisada na região do espectro de infravermelho médio compreendido entre 525 – 1800 cm^{-1}, onde, segundo Ortiz *et al.* (2013) é chamada de região da "impressão digital" do espectro, por conter as características de absorção dos principais componentes. Cada espectro apresentado na Figura 3.3 representa uma amostra que é descrita por 661 variáveis, evidenciando o elevado número de variáveis frente ao de amostras disponíveis. Para realizar os passos (3) e (4) do método proposto foi utilizado o *software MATLAB*®, versão R2012b.

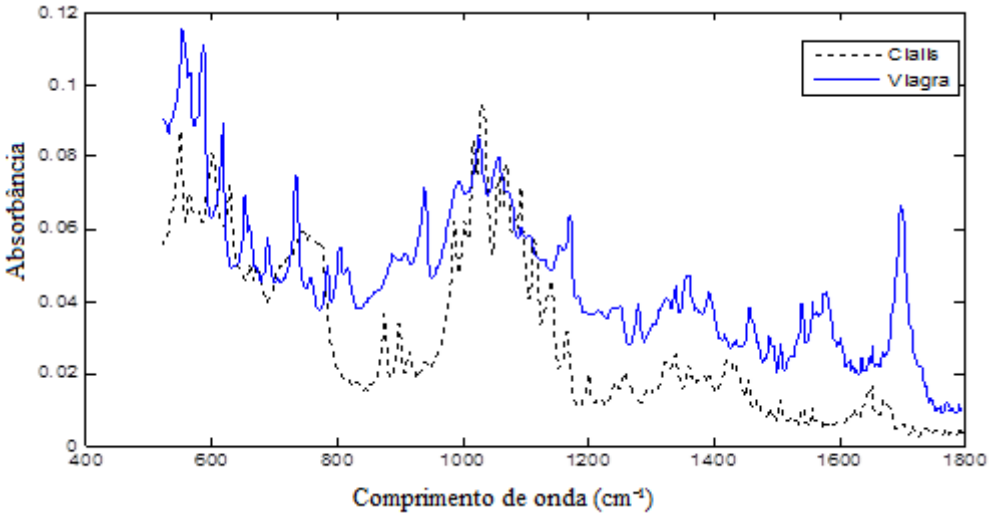

Figura 3.3: Espectros da análise ATR-FTIR do Cialis® e Viagra®

Para operacionalizar o algoritmo genético foi definida uma população inicial de 30 indivíduos, baseado em Leardi *et al*. (2002), com 70 variáveis cada, ou seja, cada cromossomo podia selecionar no máximo 70 variáveis, o que corresponde a aproximadamente 10% do total de variáveis obtidos pela análise ATR-FTIR; tal valor foi escolhido por ter apresentado mehores resultados durante os testes preliminares. As técnicas de seleção escolhidas foram o elitismo, preservando sempre os três cromossomos com melhor desempenho. A probabilidade de cruzamento e mutação foram de 80% e 1%, respectivamente, valores definidos com base nos estudos de Raymer *et al*. (2000) e Costa Filho e Poppi (2002). Por fim, para garantir a rapidez do processo devido ao elevado número de variáveis, Leardi *et al*. (2002) recomenda realizar o AG repetidas vezes com quantidades menores de gerações, de forma a variar as populações iniciais e assim conseguir testar o maior número de variáveis em menor tempo. Assim o critério de parada foi definido como 300 rodadas do AG contendo 100 gerações cada.

Testaram-se cinco valores de vizinhos mais próximos *k* (1, 3, 5, 7, 9) para classificação via *k*NN. Com o intuito de garantir que os resultados da classificação não sejam favorecidos pelo valor do parâmetro *k* ou pela divisão do conjunto de dados, foram utilizados três proporções de treino e teste, 60 - 40%, 75 - 25% e 90 - 10%. Cada classificação contou com 500 repetições e as amostras utilizadas para treino ou teste foram escolhidas

aleatoriamente em cada repetição, de forma a misturar o banco de dados. Com isso, a rotina proposta no método foi operacionalizada quinze vezes para cada banco de dados, testando assim todas as combinações dos valores de *k* com as proporções do banco de treino e teste.

A cada rodada do AG foi obtido o número de variáveis selecionadas que resultaram no maior valor de acurácia média; assim, em cada uma das quinze replicações do método foram obtidos 300 valores de acurácia e número de variáveis. Esses valores foram refinados de modo a construir um gráfico que ilustrasse o comportamento da acurácia média à medida que a quantidade de variáveis selecionadas pelo AG varia. Nos eventos em que os números de variáveis selecionadas se repetiam, optou-se pelo número de variáveis relacionado ao maior valor de acurácia média. As Figuras 3.4 e 3.5 apresentam, respectivamente, o gráfico que relaciona a acurácia média (obtida pelo *k*NN utilizando *k* igual a 1, e a proporção do banco de treino e teste de 60 – 40%) com o percentual de variáveis retidas pelo AG quando aplicado nos dados do Cialis® e Viagra®.

O ponto indicado pela seta é referente ao subconjunto de variáveis retidas que obteve a maior acurácia média (0,997 com 8 variáveis retidas, de acordo com a Tabela 3.1) para os dados do Cialis® e acurácia média de 0,9006 com 6 variáveis retidas do Viagra® (Tabela 3.2). Percebe-se que os gráficos apresentam uma retenção de menos de 5% das variáveis originais, demonstrando que os maiores valores de acurácia média obtido pelo método proposto foram alcançados com menos da metade das variáveis que os parâmetros definidos pelo AG permitiam selecionar, que era de 70 variáveis, aproximadamente 10%.

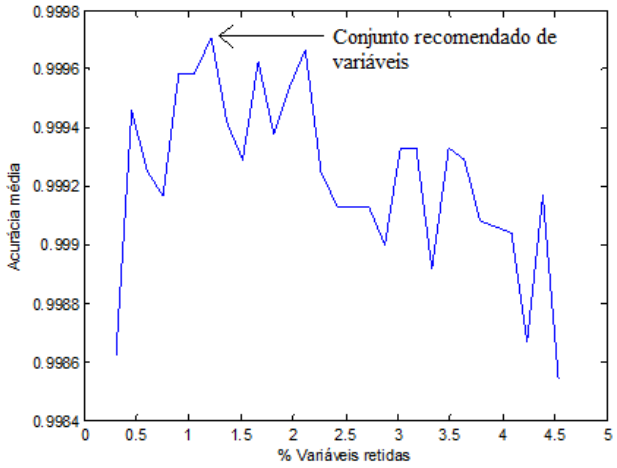

Figura 3.4: Acurácia média *vs.* porcentagem de variáveis retidas utilizando *k* = 1 e a proporção 60-40 no banco Cialis®

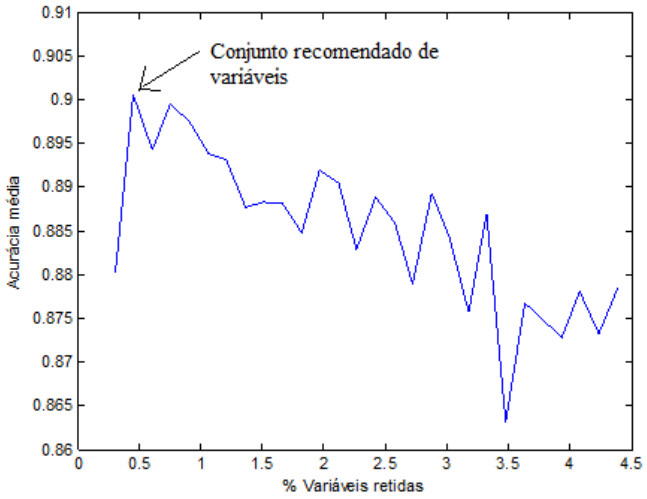

Figura 3.5: Acurácia média *vs*. porcentagem de variáveis retidas utilizando *k* = 1 e a proporção 60-40 no banco Viagra®

As Tabelas 3.1 e 3.3 apresentam os valores de acurácia média e número de variáveis retidas para distintos valores de *k* e proporções de treino e teste testados. De forma geral, percebe-se que a utilização de um subconjunto reduzido de variáveis aumenta o valor de acurácia média, reforçando a importância de eliminar variáveis ruidosas. É visto também que, à medida que a proporção do banco de treino aumenta, a acurácia aumenta, mostrando que, quanto maior a informação oferecida para construir o modelo, maior a precisão da ferramenta para a classificação. Por sua vez, aumentos no *k* sugerem prejuízos à precisão da classificação.

Para os dados do Cialis®, nota-se uma melhora de até 8% nos valores de acurácia obtidos, utilizando uma faixa de 0,3 a 1,36% das variáveis originais. A seleção de variáveis mostrou ter maior influência na acurácia quando se usa *k* igual a 9, independente da proporção do banco de treino. O melhor cenário ocorreu ao utilizar-se a proporção 75 – 25% e *k* igual a 1, obtendo acurácia = 1 com duas variáveis retidas. Nos dados do Viagra®, tem-se uma retenção de 0,3 a 1,97% das variáveis originais e um aumento de até 20% da acurácia em cada rotina. O melhor cenário ocorreu ao utilizar a proporção 90 – 10% e *k* = 1, apresentando como resultados acurácia média de 0,9386 e 5 variáveis retidas. Nas Tabelas 3.2 e 3.4 é identificado os comprimentos de onda que foram retidos em cada cenário apresentado nas Tabelas 3.1 e 3.3, repesctivamente,

Tabela 3.1 – Acurácia média e número de variáveis retidas para diferentes valores de *k* e proporções de treino e teste do banco Cialis®

		Proporção treino e teste					
		60 – 40%		75 – 25%		90 – 10%	
		Sem seleção	Com seleção	Sem seleção	Com seleção	Sem seleção	Com seleção
k=1	Acurácia média	0,9918	0,9997	0,9966	1	0,9993	1
	nº variáveis retidas	661	8	661	2	661	5
k=3	Acurácia média	0,9738	0,9982	0,9835	0,9998	0,9887	1
	nº variáveis retidas	661	9	661	8	661	2
k=5	Acurácia média	0,9440	0.9978	0,9611	0,9997	0,9730	1
	nº variáveis retidas	661	4	661	7	661	8
k=7	Acurácia média	0,9260	0.9944	0,9428	0,9991	0,9485	1
	nº variáveis retidas	661	5	661	4	661	5
k=9	Acurácia média	0,9099	0,9880	0,9255	0,9978	0,9503	1
	nº variáveis retidas	661	5	661	6	661	4

Tabela 3.2 - Identificação dos comprimentos de ondas (variáveis) retidos nos diferentes valores de *k* e proporções de treino e teste do banco Cialis®

		Proporção treino e teste		
		60 – 40%	75 – 25%	90 – 10%
k=1	Comprimentos de onda retidos	775, 800, 931, 1155, 1497, 1657, 1705, 1720	773, 1155	590, 760, 762, 1086, 1687
k=3	Comprimentos de onda retidos	769, 860, 883, 1153, 1184, 1568, 1583, 1587, 1718	771, 1155, 1377, 1385, 1537, 1589, 1637, 1705	756, 924
k=5	Comprimentos de onda retidos	771, 1151, 1585, 1689	775, 1153, 1174, 1279, 1402, 1687, 1734	773, 814, 1153, 1242, 1491, 1651, 1686, 1689
k=7	Comprimentos de onda retidos	777, 1153, 1155, 1539, 1630	773, 1149, 1537, 1768	773, 1153, 1419, 1425, 1691
k=9	Comprimentos de onda retidos	773, 1153, 1155, 1581, 1687	775, 1155, 1315, 1535, 1539, 1622	771, 1147, 1153, 1531

Tabela 3.3 – Acurácia média e número de variáveis retidas para diferentes valores de *k* e proporções de treino e teste do banco Viagra®

		Proporção treino e teste					
		60 – 40%		75 – 25%		90 – 10%	
		Sem seleção	Com seleção	Sem seleção	Com seleção	Sem seleção	Com seleção
k=1	Acurácia média	0,8338	0,9006	0,8441	0,9236	0,8523	0,9386
	nº variáveis retidas	661	3	661	3	661	5
k=3	Acurácia média	0,7755	0,8531	0,8146	0,8828	0,8329	0,9071

	nº variáveis retidas	661	3	661	8	661	13
k=5	**Acurácia média**	0,7172	0,8220	0,7693	0,8633	0,8054	0,8929
	nº variáveis retidas	661	5	661	2	661	6
k=7	**Acurácia média**	0,6642	0,7767	0,7232	0,8449	0,7649	0,8843
	nº variáveis retidas	661	2	661	2	661	3
k=9	**Acurácia média**	0,6301	0,7319	0,6592	0,7796	0,7109	0,8566
	nº variáveis retidas	661	2	661	3	661	5

Tabela 3.4 - Identificação dos comprimentos de ondas (variáveis) retidos nos diferentes valores de *k* e proporções de treino e teste do banco Viagra®

		Proporção treino e teste		
		60 – 40%	**75 – 25%**	**90 – 10%**
k=1	**Comprimentos de onda retidos**	960, 1134, 1178	660, 972, 1138	708, 947, 1124, 1151, 1747
k=3	**Comprimentos de onda retidos**	964, 1138, 1759	714, 719, 808, 958, 1007, 1105, 1120, 1371	606, 633, 679, 920, 966, 989, 1039, 1119, 1128, 1134, 1508, 1570, 1772
k=5	**Comprimentos de onda retidos**	957, 966, 980, 1138, 1140	976, 1138	704, 710, 746, 964, 1140, 1207
k=7	**Comprimentos de onda retidos**	958, 1138	960, 1140	958, 1120, 1635
k=9	**Comprimentos de onda retidos**	1126, 1759	957, 1124, 1522	860, 955, 1134, 1140, 1776

Para verificar a dispersão dos valores de acurácia média obtidos pelas iterações do GA em cada replicação do método foi gerada uma curva de distribuição da acurácia média, ilustrada nas Figuras 3.6 e 3.7. Constata-se que, para o Cialis®, as proporções 75 – 25% e 90 – 10% apresentam uma homogeneidade maior dos valores de acurácia média obtidos, apresentando curvas mais estreitas, com valores mais próximos da média e com baixo desvio padrão, apresentando resultados mais robustos quando comparada à proporção 60 – 40%. Já para o Viagra®, as proporções apresentam curvas de distribuição com dispersões similares, porém diferindo substancialmente em relação à média (pois, como visto anteriormente, à medida que a proporção do banco de teste aumenta o valor de acurácia média também aumenta).

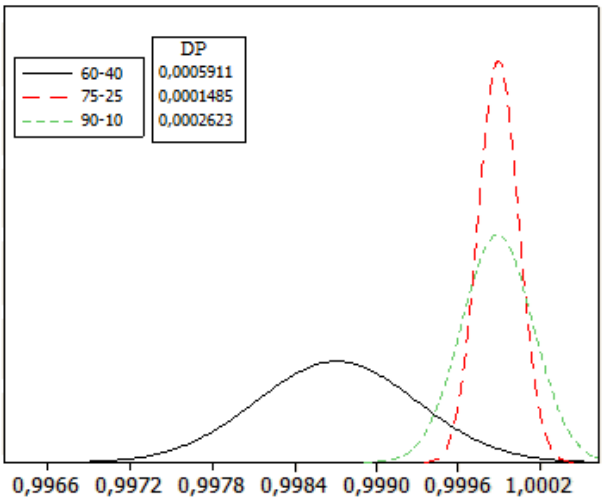

Figura 3.6: Representação da distribuição dos valores de acurácia para as três proporções de treino e teste com *k* = 1 do Cialis®

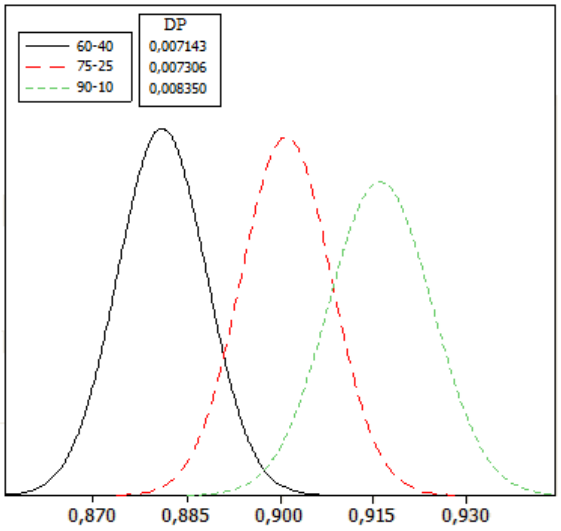

Figura 3.7: Representação da distribuição dos valores de acurácia para as três proporções de treino e teste com *k* = 1 do Viagra®

3.5 Conclusões

Neste artigo foram utilizados dados oriundos da análise espectroscópica ATR-FTIR, uma técnica bastante utilizada para a detecção de medicamentos falsificados. Dados do tipo ATR-FTIR são caracterizados por elevado número de variáveis que tendem a prejudicar o desempenho das técnicas multivariadas. Este artigo apresenta um método de seleção de variáveis para a classificação de amostras de medicamentos em originais ou falsificados. O método é apoiado no algoritmo genético para gerar subconjuntos de variáveis que serão

utilizadas na classificação das amostras via *k*NN. Cada classificação gera um valor de acurácia, que mede a precisão da classificação, e constroi-se um gráfico relacionando a acurácia média com o número de variáveis retidas para identificar o conjunto de variáveis adequado.

O método foi aplicado para cinco valores diferentes de *k* e três proporções do banco de treino e teste, em dois bancos de dados de medicamentos, Viagra® e Cialis®. Foi observado que a acurácia média aumenta em até 8% para o Cialis® e até 20% para o Viagra® quando se realiza a classificação com menos de 2% das variáveis originais. Percebeu-se ainda que, ao aumentar o tamanho do banco de treino, o desempenho do classificador melhora, em contrapartida, ao aumentar o valor de *k* a precisão do classificador diminui.

Sugere-se, para trabalhos futuros, a aplicação de outras técnicas multivariadas que possam selecionar variáveis com vistas ao aumento da precisão da identificação de medicamentos falsificados, como o *Particle Swarm Optimization*.

3.6 Referências Bibliográficas

ALADAG, C. H.; YOLCU, U.; EGRIOGLU, E.; BAS, E. Fuzzy lagged variable selection in fuzzy time series with genetic algorithms. Applied Soft Computing, v. 22, p. 465-473, 2014.

AMES, J.; SOUZA, D. Z. *Falsificação de medicamentos no Brasil*. Revista Saúde Pública, v. 46, p. 154-159, 2012.

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. *Selecting the best variables for classifying production batches into two quality levels*. Chemometrics and Intelligent Laboratories Systems, v. 97, 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 Preference, v. 22, p. 139-148, 2011.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGER, 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, p. 209-214, 2013.

BASGALUPP, M. P. *Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócio*. Dissertação de mestrado, Porto Alegre: PUC-RS, 2007.

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

COSTA FILHO, P. A.; POPPI, R. J. *Algoritmo genético em química*. Química Nova, v. 23, p. 405-411, 1999.

COSTA FILHO, P. A.; POPPI, R. J. *Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose*. Química Nova, v. 25, p. 46-52, 2002.

DECONINCK, E.; SACRÉ, P. Y.; COURSELLE, P.; DE BEER, J. O. *Chemometrics and chromatographic fingerprints to discriminate and classify counterfeit medicines containing PDE-5 inhibitors*. Talanta, v. 100, p. 123-133, 2012.

DÉGARDIN, K.; ROGGO, Y.; BEEN, F.; MARGOT, P. *Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics*. Analytica Chimica Acta, v. 705, p. 334-341, 2011.

ENGELBRECHT, A. P. *Computational Intelligence: An Introduction*. John Wiley & Sons, Ltd, West Sussex, England, 2008.

FERNANDEZ, F. M.; HOSTETLER, D.; POWELL, K.; KAUR, H.; GREEN, M.; MILDENHALL, D. C.; NEWTON, P. N. *Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries*. Analyst, v. 136, p. 3073-3082, 2011.

GOLDBERG, D.E; *Genetic Algorithms in Search, Optimization, and Machine Learnig*; Reading, Mass.: Addison-Wesley, 1998.

HAN, E. H.; KARYPIS, G.; KUMAR, V. *Text categorization using weighted adjusted K-nearest neighbor classification*. Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, p.53-65, April 16-18, 2001.

HIND, A. R.; BHARGAVA, S. K.; MCKINNON, A. *At the solid/liquid interface: FTIR/ATR – the tool of choice*. Advances in Colloid and Interface Scienc, v.93, p. 91-114, 2001.

IZQUIERDO, R. C. *Projeto de formação de células de manufatura através da utilização de algoritmos genéticos*. Dissertação de mestrado, Porto Alegre: UFRGS, 2013.

JUNG, C. R.; ORTIZ, R. S.; LIMBERGER, R.; MAYORGA, P. *A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis.* Forensic Science International, v. 216, p. 92-96, 2012.

KELLY JR., J. D.; DAVIS, L. *A Hybrid Genetic Algorithm for Classification*. International Joint Conference on Artificial Intelligence, v. 2, p. 645-650, 1991.

KIM, D. H.; ABRAHAM, A.; CHO, J. H. *A hybrid genetic algorithm and bacterial foraging approach for global optimization*. Information Sciences, v. 177, p. 3918-3937, 2007.

KONZEN, P. H. A.; FURTADO, J. C.; CARVALHO, C. W.; FERRÃO, M. F.; MOLZ, R. F.; BASSANI, I. A.; HÜNING, S. L. *Otimização de métodos de controle de qualidade de Fármacos usando algoritmo genético e busca tabu*. Pesquisa Operacional, v.23, p.189-207, 2003.

LAVINE, B. K.; DAVIDSON, C. E.; MOORES, A. J. *Innovative genetic algorithms for chemoinformatics*. Chemometrics and Intelligent Laboratory Systems, v. 60, p. 161-171, 2002.

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, p. 189-200, 2002.

LEBEL, P.; GAGNON, J.; FURTOS, A.; WALDRON, K. C.; *A rapid, quantitative liquid chromatography-mass spectrometry screening method for 71 active and 11 natural erectile dysfunction ingredients present in potentially adulterated or counterfeit products*. Journal of Chromatography A, v. 1343, p. 143-151, 2014.

LEE, Y. *Handwritten digit recognition using k nearest neighbour, radial basis function, and back-propagation neural networks*. Neural Computation, v. 3, p. 440–449, 1991.

LIU, H. H.; ONG, C. S. *Variable selection in clustering for marketing segmentation using genetic algorithms*. Expert Systems with Application, v. 34, p. 502-510, 2008.

LI, Y.; ZHANG, X. *Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection*. Chemometrics and Intelligent Laboratory Systems, v. 136, p, 47-57, 2014.

LOPES, M. B.; WOLFF, J-C. *Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging*. Analytica Chimica Acta, v. 633, p. 149-155, 2009.

MEDEIROS, A. R. B. *Uso de ATR/FTIR e FTNIR associado a técnicas quimiométricas para a quantificação de aditivos em gasolina automotiva*. Tese de Dissertação, Brasília, 2009.

ORTIZ, R. S.; MARIOTTI, K. C.; FANK, B.; LIMBERGER, R. P.; ANZANELLO, M. J.; MAYORGA, P. *Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the same pharmaceutical powder mixture be used to falsify two medicines?* Forensic Science International, v. 226, p. 282-289, 2013.

RAYMER, M. L.; PUNCH, W. F.; GOODMAN, E. D.; KUHN, L. A.; JAIN, A. K. *Dimensionality Reduction Using Genetic Algorithms*. IEEE Transaction on Evolutionary Computation, v. 4, p. 164-171, 2000.

WAN, C. H.; LEE, L. H.; RAJKUMAR, R.; ISA, D. *A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine*. Expert Systems with Applications, v. 39, p. 11880-11888, 2012.

WEISS, S. M.; APTE, C.; DAMERAU, F. J.; JOHNSON, D. E.; OLES, F. J.; GOETZ, T.; HAMPP, T. *Maximizing text-mining performance*. IEEE Intelligent Information Retrieval, v. 14, p. 63–69, 1999.

WHO; IMPACT. Counterfeit drugs kill. International medical products anti-counterfeiting taskforce, 2008. Disponível em <http://www.who.int/entity/impact/FinalBrochureWHA2008a.pdf>. Acesso em 15 de junho de 2014.

WIEGAND, P.; PELL, R.; COMAS, E. *Simultaneous variable selection and outlier detection using a robust genetic algorithm*. Chemometrics and Intelligent Laboratory Systems, v. 98, p. 108-114, 2009.

WILL, A.; BUSTOS, J.; BOCCO, M.; GOTAY, J.; LAMELAS, C. *On the use of niching genetic algorithms for variable selection in solar radiation estimation*. Renewable Energy, v. 50, p. 168-176, 2013.

XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J. W.; HOLMES, M.; HANPIN, M. *Variables selection methods in near-infrared spectroscopy.* Analytica Chimica Acta, v. 667, p. 14-32, 2010.

ZUO, W. L.; WANG, Z. Y.; LIU, T.; CHEN, H.L. *Effective detection of Parkinson's disease using an adaptive fuzzy* k-*nearest neighbor approach*. Biomedical Signal Processing and Control, v. 8, p. 364-373, 2013.

4 Terceiro Artigo: Algoritmo Genético na seleção de intervalos de variáveis ATR-FTIR com vistas à categorização de medicamentos em duas classes

Gabrielli Harumi Yamashita

Michel José Anzanello

Resumo

A análise de perfil por espectroscopia de infravermelho ATR-FTIR é normalmente aplicada em amostras de medicamentos apreendidos para detectar falsificações. Ao se trabalhar com dados espectroscópicos percebe-se a necessidade da seleção de variáveis significativas, visto que os resultados gerados pela análise ATR-FTIR possuem variáveis excessivamente ruidosas e correlacionadas. Este artigo propõe um método de seleção de variáveis espectroscópicas mais relevantes para a classificação de medicamentos em originais ou falsificados. A técnica de divisão do espectro em intervalos é inicialmente aplicada de forma a identificar as regiões que possuem as variáveis mais relevantes para a classificação via *k*NN (*k-nearest neighbor*). As regiões retidas são então refinadas através do Algoritmo Genético (AG), objetivando-se excluir as variáveis ruidosas remanescentes naqueles intervalos. A divisão por intervalos no banco de ATR-FTIR de Viagra® indicou dez regiões relevantes que conduziram a um aumento de acurácia de até 13% utilizando aproximadamente 2% das variáveis originais.

Palavras-chave: Seleção de variáveis, Divisão por intervalos, *k*NN, Algoritmo Genético.

4.1 Introdução

Medicamentos falsificados são aqueles que foram enganosamente rotulados quanto à identidade ou fonte, impossibilitando a confiabilidade nas informações sobre a origem das matérias-primas, condições de manipulação dos produtos e concentração dos ingredientes farmacológicos ativos, tornando o uso desses medicamentos nocivo à saúde (WHO, 2012; ANZANELLO *et al*., 2013). Uma gama de medicamentos possuem versões falsificadas, desde os mais comuns (como analgésicos) até os mais complexos, que podem acarretar risco de

morte. Isso se deve à facilidade de acesso dos falsificadores às tecnologias necessárias para copiar os medicamentos originais, alto custo desses medicamentos no mercado e fiscalização deficiente (WHO, 2012; FERNANDEZ *et al*., 2011 e LEBEL *et al*., 2014).

Dados da polícia federal brasileira mostram que entre janeiro de 2007 e setembro de 2010, 80% dos medicamentos falsificados apreendidos continham inibidores da fosfodiesterase tipo 5 (PDE-5) para disfunção erétil, dentre os quais destacam-se as marcas comerciais Cialis® e Viagra® (JUNG *et al*,. 2012). Esse índice elevado é resultado do sucesso desses medicamentos no mercado, da sua comercialização sem restrições pela internet e o constragimento do consumidor em adquirir esses produtos nas farmácias (HOLZGRABE *et al.*, 2011).

Para averiguar a autenticidade dos medicamentos apreendidos de forma rápida e confiável tem-se utilizado a análise de perfil por espectroscopia de infravermelho ATR-FTIR (*Attenuated Total Reflectance - Fourier Transform Infrared*), caracterizada por ser uma técnica de obtenção de espectros na região do infravermelho que requer pouco ou nenhum preparo para a maioria das amostras, além de apresentar versatilidade em termos de técnicas de amostragem (ORTIZ *et al.*, 2013; SETTLE, 1997). A ATR-FTIR tem sido cada vez mais adotada como ferramenta analítica em diversos campos, como na indústria farmacêutica, forense, alimentícia, química e petroquímica; quando combinada a técnicas multivariadas percebe-se a melhora da qualidade dos resultados obtidos (ANZANELLO *et al*., 2013; TARANTILIS *et al*., 2008; FERRÃO *et al*., 2011; FILGUEIRAS *et al*., 2014).

Entretanto, a espectroscopia no infravermelho tipicamente gera resultados com elevado número de variáveis ruidosas e altamente correlacionadas, tornando importante a aplicação de técnicas para seleção de variáveis para garantir a construção de modelos robustos de classificação ou predição (ANZANELLO *et al*., 2013). Com este propósito, métodos de seleção de variáveis por intervalos e métodos de otimização como o algoritmo genético (AG) têm sido vastamente utilizados (FERRÃO *et al*., 2011; KONZEN et al, 2003).

A seleção de variáveis por intervalos auxilia na identificação da região do espectro que contém as informações mais importantes para a geração de modelos eficazes (NORGAARD *et al*., 2000). Diversas técnicas multivariadas descritas na literatura têm sido integradas à seleção por intervalos de forma a identificar as variáveis mais relevantes contidas em cada

região do espectro (BORIN e POPPI, 2005). A seleção de variáveis apoiada no AG tem sido bastante utilizada devido à sua eficiência, robustez e versatilidade (GOLDBERG, 1998). Uma das aplicações mais difundidas do AG está na seleção de comprimentos de onda (variáveis) espectroscópicas utilizando calibração multivariada, com o objetivo de determinar o número mínimo de variáveis que possuam informação relevante para a análise (COSTA FILHO E POPPI, 1999).

Este artigo utiliza a técnica de divisão do espectro em intervalos (FERRÃO *et al*., 2011) com o propósito de identificar as regiões que contenham as variáveis mais relevantes para classificação via *k*NN de amostras de medicamentos em duas classes (autêntica e falsa). Em um segundo momento, propõe-se o refino das regiões selecionadas, que contêm as variáveis mais informativas, através do AG; tal etapa visa excluir as variáveis ruidosas pertencentes aos intervalos selecionados, restando apenas as variáveis mais relevantes em termos de potencial de classificação de amostras.

Este artigo está estruturado como segue, além desta introdução. Na seção 2 é apresentado um referencial teórico sobre seleção por intervalos, Algoritmo Genético e classificação via *k*NN, na seção 3 é descrito o método proposto. Na seção 4 são apresentados os resultados obtidos em conjunto com as discussões referentes. A seção 5 traz as considerações finais.

4.2 Referencial Teórico

4.2.1 Métodos de seleção por intervalos *i*

A seleção de variáveis por intervalos consiste em dividir o espectro em regiões equidistantes e gerar modelos em cada subintervalo, assim como em todo o espectro, de forma a comparar os resultados gerados e obter a região do espectro que apresente resultados mais consistentes (FERRÃO *et al*., 2011).

A quantidade de intervalos a ser utilizada em um experimento é definida pelo analista, porém recomenda-se cuidado com a escolha, pois ao dividir o espectro em um reduzido número de intervalos (compostos por elevado número de variáveis) pode-se fazer com que uma região com bom potencial para geração de modelos seja “contaminada” por variáveis

ruidosas; de forma complementar, dividir os dados em intervalos muito estreitos pode levar a regiões pobres em informações (NORGAARD *et al*., 2000).

Dentre as principais vantagens deste método estão a possibilidade de visualizar graficamente o modelo que está sendo investigado (facilitando a identificação dos intervalos mais relevantes), permitir a comparação entre modelos distintos no mesmo intervalo, e apresentar alta sensibilidade para identificar e excluir variáveis ruidosas (BORIN e POPPI, 2005; NORGAARD *et al*., 2000).

Em grande parte dos estudos reportados pela literatura, o método de seleção de variáveis por intervalos está integrado à regressão de mínimos quadrados parciais (*i*PLS). Suhandy *et al*. (2011) utilizam *i*PLS no desenvolvimento de modelos para medir a concentração de vitamina C em solução aquosa, enquanto que Chen *et al*. (2008) aplicam o *i*PLS na determinação do conteúdo de polifenóis totais no chá verde. Borin e Poppi (2005) fazem uso do *i*PLS para investigar e determinar quantidades de contaminantes presente no óleo lubrificante, ao passo que Ferrão *et al*. (2011) estudam os parâmetros de qualidade do biodiesel através da técnica *i*PLS.

4.2.2 Algoritmo Genético (AG) e sua utilização em seleção de variáveis

A seleção de variáveis apoiada no Algoritmo Genético (AG) tem sido vastamente utilizada devido à sua eficiência, robustez e versatilidade (GOLDBERG, 1998). O AG é um algoritmo de busca aleatória direcionada, que se baseia na seleção natural para chegar a um conjunto de variáveis que produza um resultado consistente em termos de predição ou classificação (WIEGAND *et al*., 2009). O princípio fundamental do AG é que as gerações derivadas serão mais evoluídas do que seus antecedentes; de tal forma, indivíduos melhores continuariam existindo enquanto que indivíduos mais frágeis tenderiam a ser eliminadas (KONZEN *et al.*, 2003).

A execução do AG inicia com a geração da população inicial (solução inicial) que é constituída por um número finito de indivíduos, chamado de cromossomos; cada um deles é composto por um grupo de genes codificados (KIM *et al*., 2007). O algoritmo foi estruturado de forma que as informações referentes a um determinado sistema pudessem ser codificadas de maneira análoga aos cromossomos biológicos, assemelhando-se ao processo evolutivo

natural. Quando se considera seleção de variáveis espectroscópicas, a codificação se dá de forma que cada gene do indivíduo represente uma das variáveis do espectro, fazendo com que o indivíduo contenha todas as variáveis do espectro (COSTA FILHO E POPPI, 1999). O código binário é utilizado, e cada gene pode receber o valor 1 ou 0, onde 1 representa que a variável foi selecionada e 0 representa que a variável não foi selecionada para fazer parte do modelo (RAYMER *et al*., 2000).

Cada indivíduo da população corresponde a um ponto no espaço de busca e uma possível solução para o problema em questão (KIM *et al*., 2007). Com isso, eles são avaliados através da função de aptidão, que representa a função objetivo a ser otimizada, calculando a aptidão de determinado indivíduo em relação ao modelo e avaliando se o indivíduo está apto a reproduzir e permanecer na população (GOLDBERG, 1998; KONZEN *et al.*, 2003). O desempenho, calculado pela função de aptidão, representará a chance de cada indivíduo em participar do processo reprodutivo nas próximas gerações (WHITLEY, 1994).

A evolução do AG se dá através de operadores inspirados no processo de evolução natural, conhecidos como operadores genéticos de seleção, cruzamento e mutação, que manipulam os indivíduos de uma população, através das gerações, de forma a melhorar a adaptação de cada indivíduo (KONZEN *et al*., 2003). O processo de seleção garante que os indivíduos mais adaptados tenham maiores chances de sobreviver ou reproduzir e é aplicada na população corrente para criar uma população intermediária que irá passar pelos processos de cruzamento e mutação para a geração da próxima população (WHITLEY, 1994).

Entre as técnicas de seleção mais utilizadas estão: (i) roleta, onde cada indivíduo tem a probabilidade de ser selecionado proporcional ao seu valor de aptidão, aptidão, assim os indivíduos que possuem uma alta aptidão ocuparão uma porção maior na roleta do que os indivíduos que possuem uma aptidão menor, e então a roleta é girada e é selecionado um individuo que participara do processo de geração da nova população, esta ação se repete até selecionar a quantidade desejada de indivíduos; (ii) torneio, que entre dois indivíduos aleatórios seleciona o que tiver o melhor valor em um critério a ser determinado, realiza-se quantas disputas for necessária para selecionar a quantidade de indivíduos que irá reproduzir; (ii) torneio, que entre dois indivíduos aleatórios seleciona o que tiver o melhor valor em um critério a ser determinado; e (iii) elitismo, que preserva um grupo de indivíduos mais aptos da população atual para completar a nova população (GOLDBERG, 1998).

O cruzamento consiste na manipulação do material genético (variáveis) existente na população e permite a criação de um ou mais indivíduos, resultantes do cruzamento dos indivíduos selecionados pela seleção (HAUPTY e HAUPTY, 2004). Durante a permuta do material genético entre os indivíduos haverá uma tendência de transmissão das variáveis dominantes para as gerações futuras, tornando o cruzamento responsável pela convergência para a situação de otimização desejada (COSTA FILHO E POPPI, 1999). O cruzamento pode acontecer de forma heterossexual, onde é feita a distinção do gênero de um indivíduo e apenas indivíduos de gêneros diferentes podem cruzar entre si; homossexual, onde ocorre cruzamento entre indivíduos sem distinção de gênero; e assexuado, onde ocorre a troca de genes dentro do próprio indivíduo. Porém cruzamento pela distinção de gênero aumenta a diversificação dos indivíduos, visto que, na natureza determinadas características desejáveis muitas vezes podem ser encontradas somente em um determinado gênero (COSTA FILHO E POPPI, 1999).

A mutação permite que indivíduos da nova geração sofram pequenas alterações, possuindo o papel de repor ou acrescentar um material genético inexistente na população atual, que pode ter sido perdido ou nunca ter existido em populações anteriores, permitindo assim uma possibilidade de busca maior no espaço do problema. O processo inicia-se com a escolha de um gene aleatório de um indivíduo, e então é aplicada uma taxa de probabilidade de troca deste gene por outro (KOZA, 1995). No final do processo de aplicação dos operadores genéticos, a população geralmente permanece do mesmo tamanho da população anterior (HAUPTY e HAUPTY, 2004).

Para finalizar o algoritmo é necessário definir o critério de parada, que pode ser o tempo de processamento, o resultado ótimo alcançado, o número de gerações do algoritmo, ou algum outro indicador definido pelo analista, pois enquanto o AG não satisfizer o critério de parada ele persistirá na busca do resultado ótimo (KELLY e DAVIS, 1991; COSTA FILHO E POPPI, 1999).

O AG, segundo Basgalupp (2007), pode ser empregado de duas formas na seleção de variáveis: *wrapper* e *filter*. A primeira ocorre quando é necessário introduzir um classificador para o cálculo da função de aptidão, tornando o processo mais demorado, pois cada candidato a solução precisa passar por um classificador; caso não seja necessário a indução do classificador, o método pertence ao tipo *filter*. O AG tem sido cada vez mais adotado na função *wrapper* em diversas áreas, como na indústria química, farmacêutica, energia

renovável e marketing (LEARDI *et al*., 2002; KONZEN *et al*., 2003; WILL *et al*., 2013; LIU e ONG, 2008).

A aplicação do algoritmo genético na seleção de variáveis tem sido bastante eficiente por sua capacidade de buscar vários pontos em paralelo, podendo encontrar um ótimo global e evitando cair em um ótimo local (RAYMER *et al*., 2000; LIU e ONG, 2008). Dentre suas vantagens, o AG é tido como mais resistente a se prender em ótimos locais; utiliza regras de transição probabilísticas e não determinísticas; é robusto e aplicável a uma grande variedade de problemas; e é facilmente combináveis com outras técnicas e heurísticas (GOLDBERG, 1998).

4.2.3 *k Nearest Neighbors (k*NN*)*

O *k Nearest Neighbors* (*k*NN) é uma ferramenta de classificação de observações que opera como um algoritmo de aprendizagem supervisionado, utilizando uma fase de treinamento para mapear as amostras conhecidas em regiões com maior semelhança entre suas variáveis; quando uma nova amostra dá entrada no classificador, este utiliza a distância entre essa nova amostra e as classes que foram determinadas na fase de treinamento para alocar o novo ponto à sua categoria específica (HAN *et al*., 2001).

Esta ferramenta tem sido amplamente utilizada em estudos, e a única restrição que pode interferir na sua eficiência é a necessidade de determinar um valor apropriado para o parâmetro *k*, visto que a classificação é apoiada na Distância Euclidiana de uma nova amostra em relação aos seus *k* vizinhos mais próximos, e uma nova amostra é alocada a uma classe já conhecida se a maioria dos seus *k* vizinhos pertencerem àquela classe (HAN *et al*., 2001; WAN *et al*., 2012; ANZANELLO *et al*., 2013). Assim, o valor de *k* é obtido por testes que maximizem a precisão, confiabilidade ou sensibilidade da classificação no conjunto de treinamento onde a classe de cada amostra é conhecida (ANZANELLO *et al*., 2009).

Para calcular a precisão da classificação, Anzanello *et al*., (2013) utilizam uma medida de acurácia, que consiste na média entre as classificações corretas e a quantidade total de classificações realizadas pelo *k*NN, criando um intervalo de 0 a 1, onde valores próximos a 0 indicam que a maioria das amostras foram alocadas à classe errada, enquanto que valores

próximos a 1 indicam que grande parte das amostras foram classificadas corretamente. Assim, o valor que caracteriza uma classificação totalmente exata é 1.

O *k*NN tem sido uma das ferramentas mais utilizadas para a classificação, sendo aplicada em diversas áreas, como na indústria farmacêutica, alimentícia, forense, análise de qualidade e neurociência (ANZANELLO *et al.*, 2013; ANZANELLO *et al*., 2011; LI e ZHANG, 2014; CHAOVALITWONGSE *et al*., 2007). Isso se dá devido à sua simplicidade conceitual, necessita de poucos dados de amostra para treinamento, necessidade de apenas um parâmetro *k*, e elevada disponibilidade em pacotes de *software* (HAN *et al*., 2001; ANZANELLO *et al*., 2011).

4.3 Método

Com intuito de classificar as amostras dos medicamentos apreendidos em duas classes (original ou falsificado), o método proposto para a seleção das variáveis é dividido em duas fases: (1) partição do espectro e classificação de amostras via *k*NN e (2) otimização via AG dos intervalos isolados e dos intervalos combinados. Essas fases são detalhadas a seguir.

4.3.1 Fase 1 - Partição do espectro via *k*NN

O propósito dessa fase é dividir o espectro em regiões menores equidistantes e, em seguida, desenvolver um modelo de classificação, através da ferramenta *k*NN, para cada um dos subintervalos, obtendo uma medida de acurácia (precisão da classificação) para cada região.

Cada intervalo possui variáveis diferentes, e essa técnica funciona de modo que ao selecionar um intervalo, as variáveis pertencentes a ele são selecionadas em cada amostra do banco de dados e são submetidas à classificação via *k*NN. No *k*NN, o banco de dados é dividido em treino e teste, sendo esta proporção definida pelo usuário; as amostras que irão compor os dados de treino e os dados de teste são escolhidas randomicamente pela ferramenta. Para cada subconjunto de variáveis pertencentes ao intervalo são feitas diversas classificações no *k*NN, de forma a obter resultados de diversas combinações do banco de treino e do banco de teste. Na sequência é calculada a média das acurácias obtidas em cada repetição da classificação, gerado uma acurácia média.

Assim, a região do espectro que apresentar o maior valor de acurácia média é aquela composta pelas variáveis mais relevantes, que serão selecionadas para a construção dos modelos. Essa técnica de divisão por intervalos vai ser realizada com orientação vertical, como demonstra a Figura 4.1.

A obtenção dos intervalos segue o seguinte mecanismo: primeiramente divide-se o espectro ao meio, realiza-se a classificação com base nas variáveis existentes em cada metade, e então calcula-se a acurácia para cada intervalo. Na sequência, divide-se cada metade em duas novas partes (representando 25% do espectro inicial) e repete-se o procedimento de classificação. O procedimento iterativo de geração de intervalos e classificação das amostras com base nas variáveis de cada intervalo se repete até atingir-se o número de intervalos (*i*) definido pelo usuário. Diversas aplicações práticas utilizam *i*=32 (FERRÃO *et al*., 2011).

Somente são encaminhados à Fase 2 (descrita abaixo) os intervalos que cuja acurácia é superior à acurácia obtida utilizando a totalidade das variáveis.

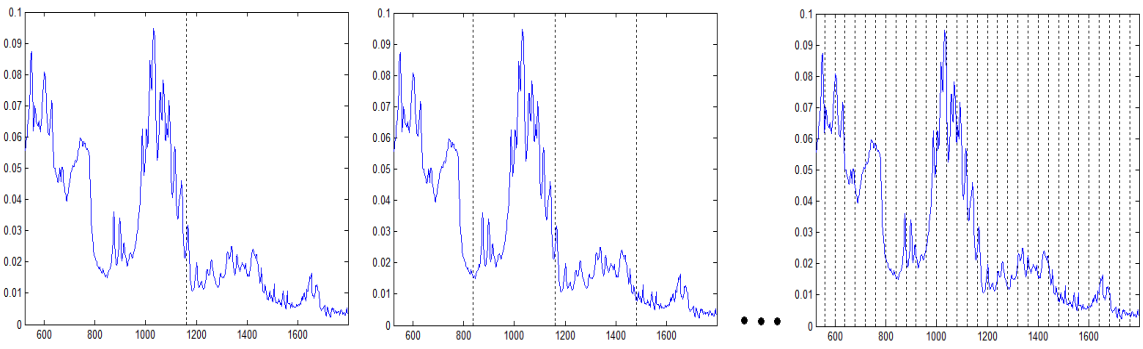

Figura 4.1: Representação da divisão do espectro em intervalos verticais

4.3.2 Fase 2 - Otimização via AG dos intervalos isolados e dos intervalos combinados

Em cada intervalo selecionado pela Fase 1 é aplicado o AG, de forma a identificar as variáveis mais relevantes pertencentes àquele intervalo e assim refinar ainda mais a seleção de variáveis. A Figura 4.2 apresenta o fluxograma da operacionalização do AG em conjunto com *k*NN. Inicialmente é gerada uma população inicial aleatória, com tamanho e quantidade de variáveis definidas de acordo com indicações da literatura. Em seguida é feita a avaliação da população, onde as variáveis pertencentes a cada indivíduo são selecionadas nas amostras e incluídas no classificador *k*NN.

A execução e configuração do *k*NN nessa etapa será igual ao utilizado na Fase 1, porém a acurácia média obtida servirá como o valor da função de aptidão do AG, sendo utilizada para avaliar a adaptação e permanência de cada indivíduo da população. Em seguida, é verificado se o critério de parada foi atendido; se sim, tem-se o melhor subconjunto de variáveis relevantes; caso contrário, é realizada a evolução dos indivíduos através da seleção, cruzamento e mutação, criando uma nova população e repetindo o ciclo.

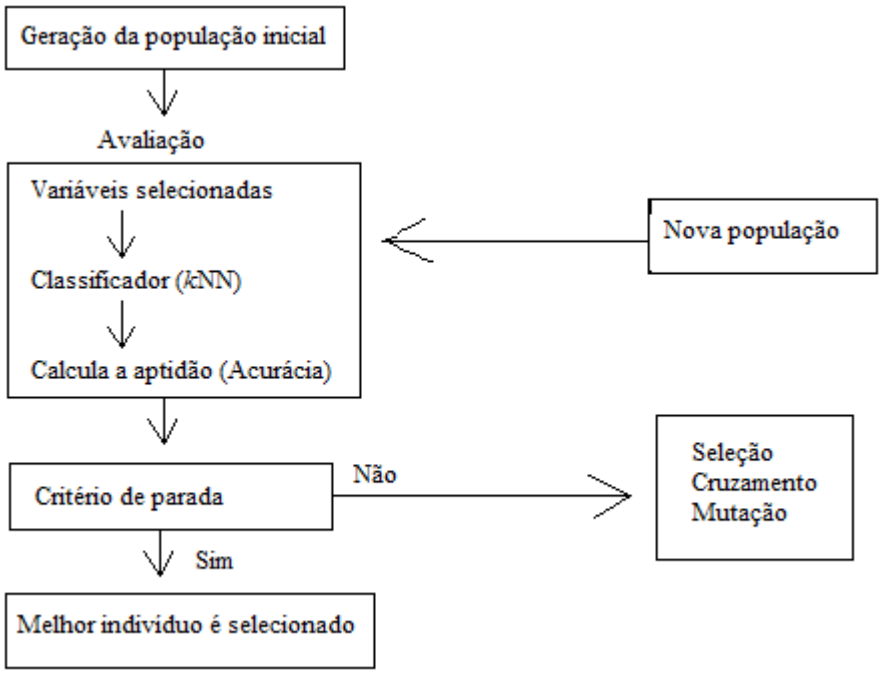

Figura 4.2: Fluxograma do *k*NN integrado ao algoritmo genético

Em seguida, este procedimento será aplicado também às combinações dos intervalos selecionados que não se sobrepõem, de modo a abranger todas as possibilidades de selecionaras variáveis que resultem na maior acurácia média.

4.4 Resultados e Discussões

Os dados utilizados neste estudo foram gerados pela análise ATR-FTIR de 69 amostras do Viagra®; sendo 21 amostras comerciais, adquiridas através dos laboratórios Pfizer Ltda. e Eli Lilly do Brasil Ltda. e em farmácias locais de Porto Alegre – RS, e 48 amostras falsificadas, fornecidas pelo departamento da Polícia Federal de Porto Alegre – RS.

Na Figura 4.3 é apresentado o resultado da análise ATR-FTIR das amostras do Viagra®, cada espectro representa uma amostra que foi analisada com base na região do espectro de infravermelho médio chamada de "impressão digital", 525 – 1800 cm^{-1}, onde, de acordo com Ortiz *et al.* (2013) ocorrem absorções características de grupos funcionais, permitindo uma melhor detecção das diferenças nos espectros. Após a análise, tem-se que cada amostra é descrita por 661 variáveis. Os procedimentos computacionais foram realizados através do *software MATLAB*®, versão R2012b.

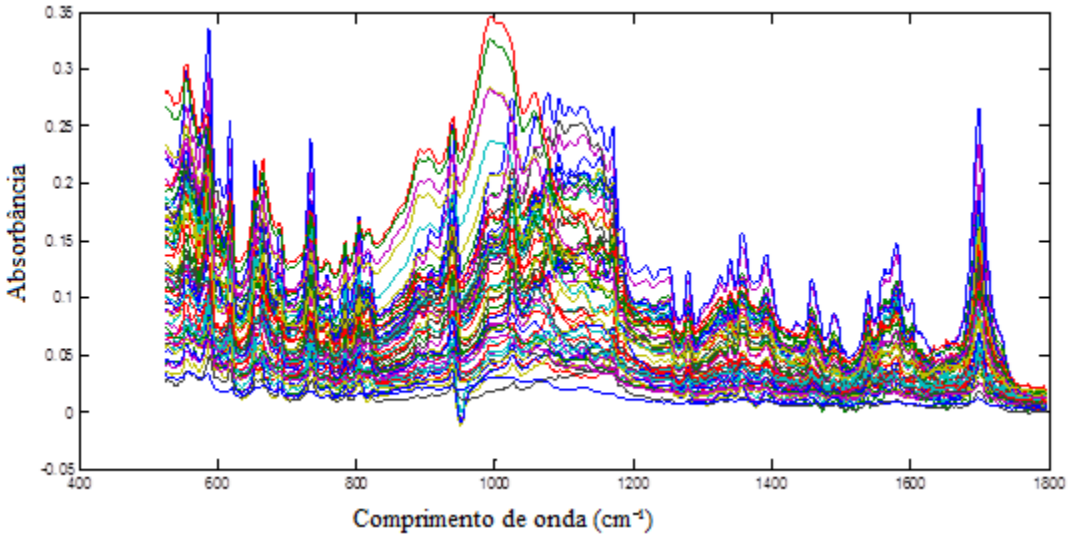

Figura 4.3: Espectros da análise ATR-FTIR das amostras do Viagra®

O espectro completo foi dividido em 2, 4, 8, 16 e 32 intervalos; a cada divisão, as variáveis pertencentes a cada intervalo foram inseridas na ferramenta de classificação *k*NN. Para o banco Viagra®, definiu-se $k = 1$, banco de treino e teste na proporção de 90 – 10%; foram realizadas 500 repetições da classificação em cada intervalo selecionado (visando alterar as observações inseridas nas porções de treino e teste), gerando um valor de acurácia média para cada divisão. Foram obtidos 62 valores de acurácia.

Ao realizar a classificação com todas as 661 variáveis, obteve-se um valor de acurácia média igual a 0,8494. Conforme propoposto na seção 4.3.1, para a realização da segunda fase do método proposto foram escolhidos apenas os intervalos responsáveis por acurácias médias superiores a 0,8494; foram obtidos 10 intervalos. Na Figura 4.4 são apresentados os intervalos de espectro selecionados para a fase 2 do método em ordem

decrescente do valor de acurácia; a Tabela 4.1 especifica a acurácia média e número de variáveis retidas em cada intervalo.

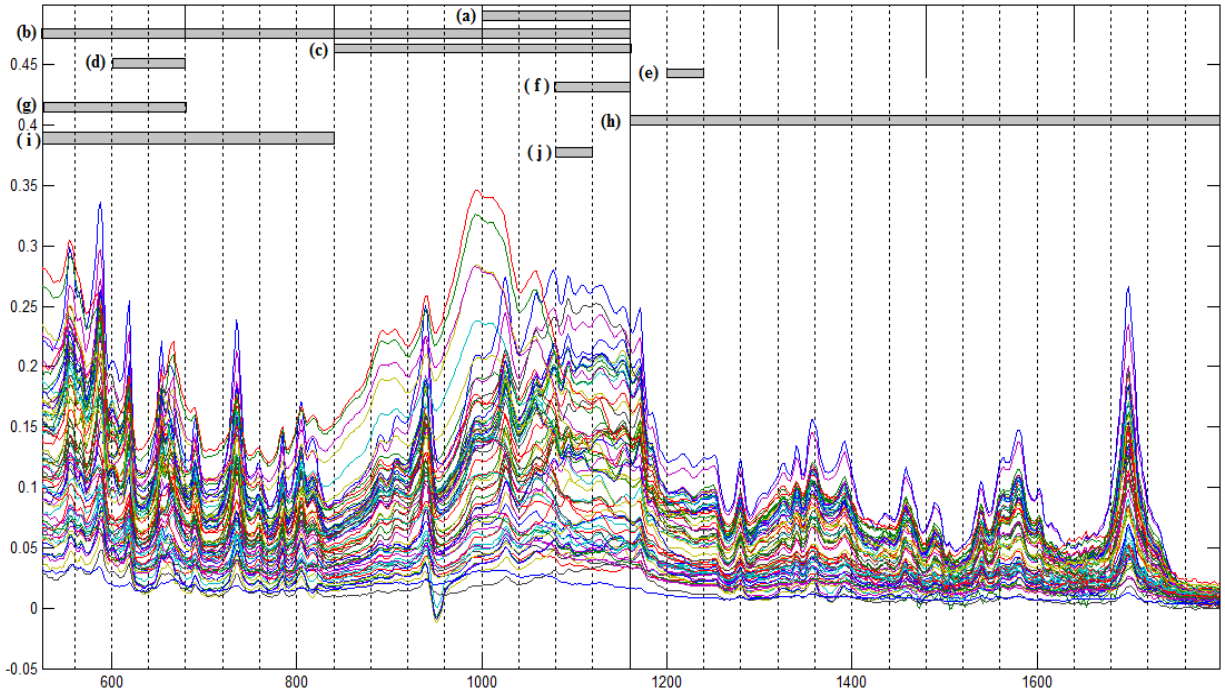

Figura 4.4: Regiões do espectro com acurácia média maior que 0,8494

Tabela 4.1– Acurácia média e número de variáveis em cada intervalo selecionado do banco Viagra®

Intervalo	Acurácia média	nº de variáveis
a	0,8991	82
b	0,8977	331
c	0,8897	165
d	0,8843	41
e	0,8691	21
f	0,8614	41
g	0,8600	83
h	0,8591	330
i	0,8569	166
j	0,8503	20

Nos dez intervalos selecionados foi aplicado o AG para refinar as variáveis a serem retidas em cada intervalo. O AG foi ajustado para operacionalizar conforme Leardi *et al.*

(2002), Raymer *et al*. (2000) e Costa Filho e Poppi (2002), com uma população inicial de 30 indivíduos, cada indivíduo composto por até 30 variáveis quando o espectro foi dividido em 2, 4 e 8 intervalos; e até 20 variáveis quando dividido em 16 e 32 intervalos. As técnicas de seleção escolhidas foram o elitismo, preservando sempre os três cromossomos com melhor desempenho, e a roleta, aplicada no restante da população. A probabilidade de cruzamento e mutação foram de 80% e 1%, respectivamente. O critério de parada foi definido após alguns testes como sendo 2000 gerações. Por fim, os parâmetros do classificador *k*NN, utilizado como função de aptidão do AG, foram os mesmos descritos anteriormente.

Ao utilizar o AG nos intervalos, percebe-se que há um aumento significativo do valor de acurácia média e diminuição do número de variáveis, justificando os esforços de refino de variáveis dentro dos intervalos. A Tabela 4.2 apresenta os valores de acurácia média e número de variáveis nos dez intervalos antes e depois da aplicação do AG. Percebe-se incremento de 4 a 8% no valor de acurácia média após a aplicação do AG; o maior valor de acurácia média pertence ao intervalo *c*, 0,9466 com apenas 16 variáveis retidas. Nota-se também que o menor valor de acurácia média obtido após o AG é maior que o maior valor de acurácia obtido nos intervalos sem a seleção via AG, reforçando a idéia da melhora que a seleção de variáveis em cada intervalo proporciona.

Tabela 4.2– Acurácia média e número de variáveis antes e depois da aplicação do AG em cada intervalo selecionado do banco Viagra®

Intervalo	Acurácia média	nº de variáveis	Acurácia média após AG	nº de variáveis após AG
a	0,8991	82	0,9446	11
b	0,8977	331	0,9431	18
c	0,8897	165	**0,9466**	**16**
d	0,8843	41	0,9300	9
e	0,8691	21	0,9086	6
f	0,8614	41	0,9120	9
g	0,86	83	0,9306	13
h	0,8591	330	0,9189	23
i	0,8569	166	0,9306	12
j	0,8503	20	0,9051	6

Na sequência foi realizado o AG nas combinações dos intervalos selecionados que não se sobrepõem (como *ad*, *ae*, *ag*, *ah*,...), resultando em 24 combinações. A Tabela 4.3 apresenta os valores de acurácia média e número de variáveis obtidos após aplicação do AG nas combinações. Percebe-se, de maneira geral, incremento na acurácia de classificação; a maior acurácia obtida refere-se ao refino do intervalo *ag*, 0,9617, retendo apenas 8 variáveis, referentes aos comprimentos de ondas 538, 555, 582, 669, 1109, 1142, 1151, 1157. Isto representa um aumento de 13% no valor de acurácia média quando comparado ao uso de todas as variáveis, valendo-se de 1,5% das variáveis originais. Percebe-se ainda que nove combinações de intervalos apresentaram valores de acurácia média superiores ao maior valor obtido com intervalos sem combinação.

Tabela 4.3 – Acurácia média e número de variáveis após aplicação do AG nas combinações dos intervalos selecionados do banco Viagra®

Combinações	Acurácia média após AG	nº de variáveis após AG	Combinações	Acurácia média após AG	nº de variáveis após AG
ad	0,9471	12	dh	0,9211	26
ae	0,9423	12	dj	0,9569	6
ag	**0,9617**	**8**	ef	0,9211	10
ah	0,9557	24	eg	0,9343	10
ai	0,9443	23	ei	0,9294	13
be	0,9449	27	ej	0,9254	11
cd	0,9466	17	fg	0,9574	15
ce	0,9477	21	fh	0,9594	29
ch	0,9409	28	fi	0,9434	10
cg	0,9437	23	gh	0,9471	29
de	0,9283	11	gj	0,9583	8
df	0,9440	12	hj	0,9471	29

4.5 Conclusões

Neste artigo é proposto um método de seleção de variáveis para a classificação de amostras, visto que, ao utilizar a análise espectroscópica ATR-FTIR para identificar a autenticidade de medicamentos, gera-se um elevado número de variáveis que tendem a

prejudicar o desempenho das técnicas multivariadas. Assim, o método é dividido em duas fases; a primeira utiliza a técnica de seleção de variáveis por intervalos, onde se divide o espectro em regiões menores realiza-se a classificação via *k*NN para cada intervalo. Os intervalos que apresentarem um valor de acurácia média maior que o valor encontrado ao realizar a classificação no espectro completo são selecionados. Na segunda fase, refinam-se os intervalos selecionados anteriormente via AG em cada intervalo e nas combinações desses intervalos que não se sobrepõem. Objetiva-se excluir as variáveis ruidosas existentes nos intervalos selecionados e identificar as variáveis mais relevantes.

Ao ser aplicado em um banco ATR-FTIR de Viagra®, a primeira etapa do método indicou dez intervalos com acurácia média maior que a acurácia do espectro completo, sendo esses selecionados para a segunda etapa do método. Na segunda fase foi observado que a acurácia média aumenta em até 8% quando comparada à seleção por intervalos e em até 13% quando comparada à classificação com todas as variáveis utilizando aproximadamente 2% das variáveis originais.

Sugere-se, para trabalhos futuros, realizar um ranqueamento dos intervalos de acordo com os valores de acurácia obtidos e realizar um processo de eliminação sistemática (*backward)* dos intervalos contendo as variáveis menos relevantes. Pode-se ainda utilizar outra técnica de classificação, como *Suport Vector Machines* (SVM).

4.6 Referências Bibliográficas

ANZANELLO, M. J.; ALBIN, S. L.; CHAOVALITWONGSE, W. *Selecting the best variables for classifying production batches into two quality levels*. Chemometrics and Intelligent Laboratories Systems, v. 97, 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 Preference, v. 22, p. 139-148, 2011.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGER, 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, p. 209-214, 2013.

BORIN, A.; POPPI, R. J. *Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil*. Vibrational Spectroscopy, v. 37, p. 27-32, 2005.

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

CHEN, Q.; ZHAO, J.; LIU, M.; CAI, J.; LIU J. *Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms*. Journal of Pharmaceutical and Biomedical Analysis, v. 46, p. 568-573, 2008

COSTA FILHO, P. A.; POPPI, R. J. *Algoritmo genético em química*. Química Nova, v. 23, p. 405-411, 1999.

FERNANDEZ, F. M.; HOSTETLER, D.; POWELL, K.; KAUR, H.; GREEN, M.; MILDENHALL, D. C.; NEWTON, P. N. *Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries*. Analyst, v. 136, p. 3073-3082, 2011.

FERRÃO, M. F.; VIEIRA, M. S.; PAZOS, R. E. P.; FACHINI, D.; GERBASE, A. E.; MANDER, L. *Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions*. Fuel, v. 90, p. 701-706, 2011.

FILGUEIAS, P. R.; SAD, C. M, S.; LOUREIRO; SANTOS, M. F. P.; CASTRO, E. V. R.; DIAS, J. C. M.; POPPI, R. J. *Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration*. Fuel, v. 116, p. 123-130, 2014.

GOLDBERG, D.E; *Genetic Algorithms in Search, Optimization, and Machine Learnig*; Reading, Mass.: Addison-Wesley, 1998.

HAN, E. H.; KARYPIS, G.; KUMAR, V. *Text categorization using weighted adjusted K-nearest neighbor classification*. Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, p.53-65, April 16-18, 2001.

HAUPTY, R. L.; HAUPTY, S.E. *Practical genetic algorithm.* 2º ed. A John Wiley & Sons,Inc., Publication, 2004.

HOLZGRABE, U.; MALET-MARTINO, M. *Analytical challenges in drug counterfeit-ing and falsification—the NMR approach*. Journal of Pharmaceutical and Biomedical Analysis, v. 55, p. 679-687, 2011.

JUNG, C. R.; ORTIZ, R. S.; LIMBERGER, R.; MAYORGA, P. *A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis.* Forensic Science International, v. 216, p. 92-96, 2012.

KIM, D. H.; ABRAHAM, A.; CHO, J. H. *A hybrid genetic algorithm and bacterial foraging approach for global optimization*. Information Sciences, v. 177, p. 3918-3937, 2007.

KONZEN, P. H. A.; FURTADO, J. C.; CARVALHO, C. W.; FERRÃO, M. F.; MOLZ, R. F.; BASSANI, I. A.; HÜNING, S. L. *Otimização de métodos de controle de qualidade de Fármacos usando algoritmo genético e busca tabu*. Pesquisa Operacional, v.23, p.189-207, 2003.

KOZA, J.R. *Survey of genetic algorithms and genetic programming.* Proc. Of Wescon 95, IEEE Press, p. 589-594, 1995.

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, p. 189-200, 2002.

LEBEL, P.; GAGNON, J.; FURTOS, A.; WALDRON, K. C.; *A rapid, quantitative liquid chromatography-mass spectrometry screening method for 71 active and 11 natural erectile dysfunction ingredients present in potentially adulterated or counterfeit products*. Journal of Chromatography A, v. 1343, p. 143-151, 2014.

LI, Y.; ZHANG, X. *Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection*. Chemometrics and Intelligent Laboratory Systems, v. 136, p, 47-57, 2014.

LIU, H. H.; ONG, C. S. *Variable selection in clustering for marketing segmentation using genetic algorithms*. Expert Systems with Application, v. 34, p. 502-510, 2008.

NORGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN, J. P.; MUNCK, L.; ENGELSEN, S. B. *Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy*. Applied Spectroscopy, v. 54, p. 413-419, 2000.

ORTIZ, R. S.; MARIOTTI, K. C.; FANK, B.; LIMBERGER, R. P.; ANZANELLO, M. J.; MAYORGA, P. *Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the same pharmaceutical powder mixture be used to falsify two medicines?* Forensic Science International, v. 226, p. 282-289, 2013.

RAYMER, M. L.; PUNCH, W. F.; GOODMAN, E. D.; KUHN, L. A.; JAIN, A. K. *Dimensionality Reduction Using Genetic Algorithms*. IEEE Transaction on Evolutionary Computation, v. 4, p. 164-171, 2000.

SETTLE, F. A. *Handbook of instrumental techniques for analytical chemistry*. Upper Saddle River: Prentice Hall, 1997.

SUHANDY, D.; YULIA, M.; OGAWA, Y.; KONDO, N. *Prediction of vitamin C using FTIR-ATR terahertz spectroscopy combined with subinterval partial least squares (iPLS) regression*. System Integration (SII), p 202-206, 2011.

TARANTILIS, P. A.; TROIANOU, V. E.; PAPPAS, C. S.; KOTSERIDIS, Y. S.; POLISSIOU, M. G. *Differentiation of Greek red wines on the basis of grape variety using*

attenuated total reflectance Fourier transform infrared spectroscopy. Food Chemistry, v. 111, p. 192–196, 2008.

WAN, C. H.; LEE, L. H.; RAJKUMAR, R.; ISA, D. *A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine*. Expert Systems with Applications, v. 39, p. 11880-11888, 2012.

WHITLEY, D. *A genetic algorithm tutorial. Springer Science + Business Media* B.V., Formerly Kluwer Academic, p. 65-85, 1994.

WHO Media centre, *Medicines: spurious/falsely-labelled/ falsified/counterfeit (SFFC) medicines.* (Fact sheet nº 275), 2012. Disponível em < http://www.who.int/mediacentre/factsheets/fs275/en/index.html>. Acesso em 30 de agosto de 2013.

WIEGAND, P.; PELL, R.; COMAS, E. *Simultaneous variable selection and outlier detection using a robust genetic algorithm*. Chemometrics and Intelligent Laboratory Systems, v. 98, p. 108-114, 2009.

WILL, A.; BUSTOS, J.; BOCCO, M.; GOTAY, J.; LAMELAS, C. *On the use of niching genetic algorithms for variable selection in solar radiation estimation*. Renewable Energy, v. 50, p. 168-176, 2013.

5 Considerações Finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

5.1 Conclusões

Esta dissertação teve como objetivo principal o desenvolvimento de sistemáticas de seleção de variáveis com vistas à clusterização e classificação de amostras de medicamentos, visto que, ao utilizar a análise de perfil por espectroscopia de infravermelho ATR-FTIR para a detecção da autenticidade de medicamentos, é obtido um elevado número de variáveis ruidosas e correlacionadas.

Com a revisão bibliográfica realizada, objetivos específicos foram definidos. São eles: (*i*) selecionar as variáveis ATR-FTIR mais relevantes para clusterização de amostras de medicamentos através de um índice de importância gerado por meio de parâmetros oriundos da ACP; (*ii*) criar Índices de Importância de Variáveis apoiados na ACP que conduzam uma remoção ordenada de variáveis; (*iii*) aplicar o AG para selecionar os subconjuntos de variáveis ATR-FTIR mais relevantes para a classificação de amostras de medicamentos; e (*iv*) utilizar a técnica de divisão do espectro em intervalos para identificar as regiões mais relevantes para classificação de amostras de medicamentos em duas classes.

Os objetivos (*i*) e (*ii*) foram atingidos no primeiro artigo, o qual apresentou um método que utiliza os parâmetros resultantes da aplicação da ACP no banco de dados do Cialis® e Viagra® para construir três índices de importância de variáveis; tais índices constituem-se na base para remoção ordenada de variáveis através de um procedimento do tipo *backward*. À cada variável removida, realiza-se a clusterização das amostras e calcula-se um SI que indica a qualidade dos agrupamentos. Observou-se que os índices apoiados no parâmetro variância explicada geram agrupamentos mais consistentes e retêm menor número de variáveis quando aplicados aos dois bancos de dados.

O objetivo (*iii*) foi alcançado no segundo artigo, que propôs um método onde subconjuntos de variáveis são gerados pelo AG e então as amostras são classificadas repetidas vezes através do *k*NN. O método foi aplicado utilizando cinco valores de *k* e três proporções de teste e treino, revelando que, ao diminuir o valor de *k* e aumentar o tamanho do banco de

treino, o desempenho do classificador melhora em até 8% para o Cialis® e em até 20% para o Viagra®; em ambos os casos, a classificação reteve menos de 2% das variáveis originais.

O objetivo (*iv*) foi atingido na primeira fase do método proposto no terceiro artigo, que consistiu em dividir o espectro em até 32 intervalos e realizar a classificação via *k*NN em cada intervalo. Deste procedimento foram selecionadas as regiões com acurácia média maior que a obtida quando da classificação utilizando o espectro completo. Os dez intervalos selecionados na primeira fase do método foram utilizados na segunda fase, aplicando o AG nos intervalos e nas combinações de intervalos que não apresentam sobreposição. Como resultado, obteve-se um aumento de até 13% no valor de acurácia média quando comparado ao uso de todas as variáveis, utilizando-se aproximadamente 1,5% das variáveis originais.

Dos métodos abordados nos três artigos desta dissertação percebe-se que o método que combina a divisão do espectro em intervalos com o AG foi o que mais se destacou, pois ao utilizar a divisão do espectro como um "filtro" que faz a pré-seleção das variáveis que serão inseridas no AG obteve-se melhores valores de acurácia média do que nos artigos anteriores. Assim este método pode trazer resultados mais consistentes em cenários onde há necessidade de estabelecer as variáveis que mais influênciam na identificação dos medicamentos falsificados.

5.2 Sugestões para trabalhos futuros

Como extensões das proposições apresentadas nessa dissertação, sugerem-se as seguintes pesquisas futuras:

a) Propor e testar novos índices de importância das variáveis apoiadas em outras técnicas multivariadas para a eliminação sistemática de variáveis com vistas à clusterização e classificação;

b) Selecionar variáveis relevantes para a clusterização das amostras falsificadas dos medicamentos de forma a detectar similaridades entre as fontes falsificadoras;

c) Utilizar o algoritmo genético para selecionar variáveis relevantes para a clusterização; e

d) Combinar outras técnicas multivariadas para a seleção de variáveis com vistas à classificação, como método *Particle Swarm Optimization*.