UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA
DEPARTAMENTO DE QUÍMICA ANALÍTICA

Aplicação de máquinas de vetores de suporte para desenvolvimento de modelos de classificação e calibração multivariada em espectroscopia no infravermelho

TESE DE DOUTORADO

Candidato: Danilo Althmann Maretto
Orientador: Prof. Dr. Ronei Jesus Poppi

Campinas, 2011

M335a

Maretto, Danilo Althmann.
Aplicação de máquinas de vetores de suporte para desenvolvimento de modelos de classificação e calibração multivariada em espectroscopia no infravermelho / Danilo Althmann Maretto. -- Campinas, SP: [s.n], 2011.

Orientador: Prof. Dr. Ronei Jesus Poppi.

Doutorado - Universidade Estadual de Campinas, Instituto de Química.

1. Máquinas de vetores de suporte. 2. Calibração multivariada. 3. Espectroscopia no infravermelho. 4. Classificação de amostras. I. Ronei, Jesus Poppi. II. Universidade Estadual de Campinas. Instituto de Química. III. Título.

Título em inglês: Application of support vector machines in development of classification and multivariate calibration models in infrared spectroscopy

Palavras-chaves em inglês: Support vector machine (SVM), Multivariate calibration, Infrared spectroscopy, Sample classification

Área de concentração: Química Analítica

Titulação: Doutor em Ciências

Banca examinadora: Prof. Dr. Ronei Jesus Poppi (orientador), Profa. Dra. Alessandra Borin (FQ-PUCCAMP), Profa. Dra. Patrícia Valderrama (UTFPR), Prof. Dr. Francisco Benedito Teixeira Pessine (IQ-UNICAMP), Prof. Dr. Jarbas José Rodrigues Rohwedder (IQ-UNICAMP)

Data de defesa: 15/07/2011

“Instruments register only through things they're designed to register. Space still contains infinite unknowns.”

Mr. Spock

“It’s a long way to the top if you wanna rock’n’roll”

Bon Scott

Agradecimentos

Ao Prof. Dr. Ronei Jesus Poppi pela oportunidade, paciência, orientação e amizade.

Às Professoras Carmen Sílvia Passos e Cecília Amélia Fazzio Escanhoela da Faculdade de Medicina da Unicamp pelas amostras, material bibliográfico e apoio.

Ao Paulo A. da Costa Filho e empresa Rhodia da França, pelas amostras e dados cedidos.

À Esalq/USP de Piracicaba pelas amostras de solos cedidos.

À Daniela, Camila e professora Laura Otoboni do CBMEG da Unicamp, pela colaboração, apoio e amizade.

À CPRM - Serviço Geológico do Brasil, em especial à Magda Pinto, por apoiar a finalização desta tese.

Aos membros do grupo LAQQA pelo apoio e amizade: Alessandra, Gilmare, Luciana, Jez, Luiz, Patacca, Marcello, Patrícia, Paulo Henrique, Thiago, Renato, Marcos, Diórginis, Werickson, Waldomiro, Guilherme, André, Márcia, Mónica, Laura.

À Bel da CPG por todo apoio e amizade.

Aos amigos do grupo LEEDS: Renata, Rafa, Dani, Bárbara, Arnaldo, Cecília, Luciana.

À todos os professores e funcionários do Instituto de Química que direta ou indiretamente contribuíram para a realização deste trabalho.

Aos grandes amigos da Panela: Almir, Ana, Indaia, Thais, Dudu, Viana, Fer, Walter, Américo e Kátia.

Amigos e colegas de banda: Hugo, Richard, Max, Dario e Jr. “The thing you hate the most, Caffeine!!!”

Aos grandes amigos: Rita (companheira de IQ), Letícia e Karininha (primas), Fer2 e Helder (headbenzi), Will e Joy (chucos), Dani e Marlon (amizades proporcionais às distâncias), Rogério (fiz uma gentileza), Nina, Dolly, Annie e Tiger (peludos).

À Fran, por todo apoio e amor.

E principalmente à minha família, a qual sempre me apoiou e sem a qual nada seria possível.

Currículo

Dados Pessoais

Nome: Danilo Althmann Maretto

Nascimento: 22/02/1979 - Campinas/SP - Brasil

Formação Acadêmica/Titulação

- Doutorado em Ciências – Área de concentração: Química Analítica (2007 – 2011). Instituto de Química, Unicamp, Campinas, Brasil.
- Mestrado em Química (2005 – 2007), Instituto de Química, Unicamp, Campinas, Brasil.
- Graduação em Química (1998 – 2004). Instituto de Química, Unicamp, Campinas, Brasil.
- Ensino profissional de nível técnico em Bioquímica (1995 – 1997), Escola Técnica Estadual "Conselheiro Antônio Prado", Campinas, Brasil

Atuação professional

- Químico Analista (junho de 2010 – Presente). CPRM - Serviço Geológico do Brasil, Belo Horizonte, Brasil
- Técnico Químico (março de 2004 – junho de 2010). Instituto de Química, Unicamp, Campinas, Brasil
- Auxiliar técnico de biologia molecular (julho de 2003 – março de 2004). Alellyx Applied Genomics, Campinas, Brasil
- Estagiário (janeiro de 2001 – junho de 2003). Cipoi - Unicamp, Campinas, Brasil
- Estagiário (março de 1999 – dezembro de 2000). CBMEG - Unicamp, Campinas, Brasil
- Técnico de Laboratório (maio de 1998 – novembro de 1998). Coopers Brasil Ltda, Campinas, Brasil

Artigos Publicados

- Carlos, C., Maretto, D.A., Poppi , R.J., Sato, M.I.Z., Maria Inês Z. Sato, C. Ottoboni, L.M.M., *Fourier transform infrared microspectroscopy as a bacterial source tracking tool to discriminate fecal E. coli strains*, Microchemical Journal, v. 99, p. 15–19, 2011.
- Romão, W., Franco, M. F., Iglesias, A. H., Sanvido, G. B., Maretto, D. A., Gozzo, F. C., Poppi, R. J., Eberlin, Marcos N., De Paoli, M. A.*Fingerprinting of bottle-grade poly(ethylene terephthalate) via matrix-assisted laser desorption/ionization mass spectrometry.* Polymer Degradation and Stability, v.95, p.666 - 671, 2010.
- Ribeiro, D. A., Maretto, D. A., Nogueira, F. C. S., Silva, M. J., Campos, F. A. P., Domont, G. B., Poppi, R. J., Ottoboni, L. M. M. *Heat and phosphate starvation effects on the proteome, morphology and chemical composition of the biomining bacteria Acidithiobacillus ferrooxidans*. World Journal of Microbiology & Biotechnology., 27(6), p. 1469-1479, 2010.
- Sussulini, A., Prado, A., Maretto, D. A., Poppi, R. J., Tasic, L., Banzato, C. E. M., Arruda, M. A. Z.,*Metabolic Profiling of Human Blood Serum from Treated Patients with Bipolar Disorder Employing H NMR Spectroscopy and Chemometrics.* Analytical Chemistry , v.81, p.9755 - 9763, 2009.
- Maretto, D. A., Mello, C., Poppi, R. J.,*Least-squares support vector machines to correct temperature-induced spectral variation in multivariate calibration*. Journal of Near Infrared Spectroscopy. , v.16, p.249 - , 2008.
- Ferrão, M. F., Mello, C., Borin, A., Maretto, D. A., Poppi, R. J. *LS-SVM: Uma nova ferramenta quimiométrica para regressão multivariada. Comparação de modelos de regressão LS-SVM e PLS na quantificação de adulterantes em leite em pó empregando NIR.* Química Nova. , v.30, p.852 - 859, 2007.
- Borin, A., Mello, C., Ferrão, M. F., Maretto, D. A., Poppi, R. J. *Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk.* Analytica Chimica Acta. , v.579, p.25 - 32, 2006.

Resumo

"APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE PARA DESENVOLVIMENTO DE MODELOS DE CLASSIFICAÇÃO E CALIBRAÇÃO MULTIVARIADA EM ESPECTROSCOPIA NO INFRAVERMELHO"

Autor: Danilo Althmann Maretto

Orientador: Ronei Jesus Poppi

O objetivo desta tese de doutorado foi de utilizar o algoritmo Máquinas de Vetores de Suporte (SVM) em problemas de classificação e calibração, onde algoritmos mais tradicionais (SIMCA e PLS, respectivamente) encontram problemas. Foram realizadas quatro aplicações utilizando dados de espectroscopia no infravermelho. Na primeira o SVM se mostrou ser uma ferramenta mais indicada para a determinação de Carbono e Nitrogênio em solo por NIR, quando estes elementos estão em solos sem que se saiba se há ou não a presença do mineral gipsita, obtendo concentrações desses elementos com erros consideravelmente menores do que a previsão feita pelo PLS. Na determinação da concentração de um mineral em polímero por NIR, que foi a segunda aplicação, o PLS conseguiu previsões com erros aceitáveis, entretanto, através da análise do teste F e o gráfico de erros absolutos das previsões, foi possível concluir que o modelo SVM conseguiu chegar a um modelo mais ajustado. Na terceira aplicação, que consistiu na classificação de bactérias quanto às condições de crescimento (temperaturas 30 ou 40°C e na presença ou ausência de fosfato) por MIR, o SIMCA não foi capaz de classificar corretamente a grande maioria das amostras enquanto o SVM produziu apenas uma previsão errada. E por fim, na última aplicação, que foi a diferenciação de nódulos cirróticos e de hepatocarcinoma por microespectroscopia MIR, a taxa das previsões corretas para os conjuntos de validação do SVM foram maiores do que do SIMCA. Nas quatro aplicações o SVM produziu resultados melhores do que o SIMCA e o PLS, mostrando que pode ser uma alternativa aos métodos mais tradicionais de classificação e calibração multivariada.

Abstract

"APPLICATION OF SUPPORT VECTOR MACHINES IN DEVELOPMENT OF CLASSIFICATION AND MULTIVARIATE CALIBRATION MODELS IN INFRARED SPECTROSCOPY"

Author: Danilo Althmann Maretto

Adviser: Ronei Jesus Poppi

The objective of this thesis was to use the algorithm Support Vector Machines (SVM) in problems of classification and calibration, where more traditional algorithms (SIMCA and PLS, respectively) present problems. Four applications were developed using data for infrared spectra. In the first one, the SVM proved to be a most suitable tool for determination of carbon and nitrogen in soil by NIR, when these elements are in soils without knowledge whether or not the presence of the gypsum mineral, obtaining concentrations of these elements with errors considerably smaller than the estimated by the PLS. In the determination of the concentration of a mineral in a polymer by NIR, which was the second application, the PLS presented predictions with acceptable errors, however, by examining the F test and observing absolute errors of predictions, it was concluded that the SVM was able to reach a more adjusted model. In the third application, classification of bacteria on the different growth conditions (temperatures 30 or 40 ° C and in the presence or absence of phosphate) by MIR, the SIMCA was not able to correctly classify the majority of the samples while the SVM produced only one false prediction. Finally, in the last application, which was the differentiation of cirrhotic nodules and Hepatocellular carcinoma by infrared microspectroscopy, the rate of correct predictions for the validation of sets of SVM was higher than the SIMCA. In the four applications SVM produced better results than SIMCA and PLS, showing that it can be an alternative to the traditional algorithms for classification and multivariate calibration.

Índice

Lista de Tabelas

Lista de Figuras

Prefácio

Prefácio

A utilização de quimiometria para extração de informações quantitativas ou qualitativas de dados químicos está sendo cada vez mais empregada, especialmente para dados obtidos através de espectroscopia na região do infravermelho, já que esse tipo de técnica analítica fornece análises simples, rápidas e não destrutivas sendo, em muitos casos, uma boa alternativa às técnicas clássicas de análise.

Do ponto de vista tanto da aplicação, quanto o da instrumentação, o espectro no infravermelho pode ser dividido em três regiões (Tabela 1), sendo estas: a região do infravermelho próximo (NIR), a região do infravermelho médio (MIR) e a região do infravermelho distante (FIR) [1]. A maior parte das aplicações tem sido realizada no infravermelho médio e próximo, amplamente utilizadas para análises qualitativas e quantitativas [2,3].

Tabela 1. Regiões espectrais do infravermelho.

Região do infravermelho	Intervalo de números de onda (ν) – (cm^{-1})	Intervalo de comprimentos de onda (λ) – (nm)
Próximo (NIR)	12800 a 4000	780 a 2500
Médio (MIR)	4000 a 200	2500 a 5000
Distante (FIR)	200 a 10	5000 a 100000

Na região do NIR as principais aplicações encontram-se na análise quantitativa de materiais industriais e agrícolas e no controle de processos, destacando também as aplicações farmacêuticas, alimentícias e petroquímicas, sendo também uma ferramenta valiosa para a identificação e determinação de aminas primárias e secundárias na presença de aminas terciárias em misturas [4].

A região do MIR é provavelmente onde se encontra a maioria das pesquisas desenvolvidas e o maior número de aplicações. Ainda hoje, a maioria

das aplicações consiste na identificação de compostos orgânicos, pois nessa região ocorrem essencialmente transições fundamentais e existe uma faixa espectral, conhecida como região de impressão digital (1.200 a 700 cm^{-1}), onde pequenas diferenças na estrutura e na constituição de uma molécula resultam em mudanças significativas na distribuição das bandas de absorção. Em consequência, uma semelhança estreita entre dois espectros nesta região, bem como nas outras, constitui forte evidência da identidade dos compostos que produziram os espectros [1,4].

Dependendo do ambiente operacional, instrumentos podem ser acoplados com acessórios de manipulação de amostra e programas computacionais para aplicações qualitativas e quantitativas. Nos dias atuais, os fabricantes de instrumentos estão desenvolvendo instrumentos cada vez mais compactos e de custo menor, sendo razoável prever dentro de um futuro próximo a miniaturização dos espectrômetros na região do infravermelho.

Em geral, na obtenção de informações qualitativas e quantitativas a partir de espectros complexos nesta região, por tratar de dados multivariados, é imprescindível a utilização de métodos quimiométricos. Tais métodos são desenvolvidos e disponibilizados em programas computacionais e são, juntamente com os avanços tecnológicos dos instrumentos, os responsáveis pela popularização do uso da espectroscopia vibracional [1].

Mais recentemente, têm surgido novos algoritmos que vem sendo testados em aplicações onde os métodos quimiométricos mais tradicionais não produzem resultados satisfatórios e dentre esses, grande atenção tem sido dada as Máquinas de Vetores de Suporte (SVM, do inglês, *Support Vector Machines*). Esses algoritmos têm grande habilidade de generalização, podem ser utilizados em sistemas não lineares e tem solução única, tornando-os muito atraentes para tratamento de dados químicos complexos.

O objetivo desta tese de doutorado foi o de utilizar o algoritmo Máquinas de Vetores de Suporte (SVM) em problemas de classificação e calibração multivariada em dados adquiridos por espectroscopia na região do infravermelho, para testar seu desempenho em relação aos obtidos por métodos quimiométricos

mais tradicionais, no caso o PLS (do inglês *Partial Least Squares*, ou Mínimos Quadrados Parciais) para os problemas de calibração multivariada e o SIMCA (*Soft Independent Modelling of Class Analogy*) para os problemas de classificação de amostras.

Os problemas de calibração aqui estudados utilizaram como amostras de estudo dois tipos de solo (contendo e não contendo o mineral gipsita) em uma aplicação, na qual foram determinadas concentrações de Carbono e Nitrogênio e polímeros com um mineral adsorvido em outra, onde foi determinada a concentração deste mineral. O SVM foi proposto como alternativa ao PLS, na tentativa de construir modelos que fornecessem menores erros médios quadráticos de previsão.

Como matriz para os problemas de classificação foram usadas amostras de origem biológica (neste caso, bactérias e tecido de fígado humano) que costumam gerar espectros complexos e com difícil distinção entre amostras diferentes. Para esses problemas foi empregada uma rotina que utiliza algoritmo genético a fim de encontrar os melhores parâmetros para aperfeiçoar o SVM.

Na aplicação utilizando as bactérias *A. ferrooxidans* a ideia foi conseguir uma classificação das mesmas quanto às condições de crescimento. Foram realizados dois experimentos: no primeiro havia diferenças quanto à temperatura de crescimento (um conjunto de colônias cresceu à 30 e outro à 40ºC) e no outro as colônias cresciam na presença ou ausência de fosfato. O trabalho utilizando o tecido de fígado humano visou a diferenciação das amostras que continham nódulos cirróticos e nódulos de hepatocarcinoma. Isso é importante para o diagnóstico e tratamento das doenças, sendo atualmente feita por microscopia convencional.

Em ambos os estudos de classificação de amostras foram construídos modelos usando SIMCA e SVM. Os resultados foram comparados através do número de acertos da classe das amostras de validação.

A apresentação deste trabalho foi dividida da seguinte forma: dois capítulos introdutórios; um tratando sobre quimiometria, calibração multivariada, pré-processamentos dos dados e algoritmos genéticos, e outro sobre o SVM; quatro

capítulos onde são discutidas as aplicações, sendo que cada um traz uma breve introdução sobre a amostra em questão, seguido do procedimento experimental e da apresentação, discussão dos resultados e conclusões do capítulo; e a finalização da tese, que se dá nas conclusões gerais do trabalho e no índice de referências bibliográficas.

Capítulo I

1. Quimiometria

Nos anos 70 surgiu dentro da química analítica uma área de pesquisa que buscava extrair de uma grande quantidade de dados químicos complexos resultados analíticos interpretáveis. O termo Quimiometria é utilizado hoje para a análise de dados, com finalidade específica dentro de um estudo químico como a otimização de um processo, classificação de dados, modelagem e monitoramento de processos multivariados, construção de modelos de regressão e desenvolvimento de inteligência artificial, entre outros [5]. Dentro da Quimiometria, as duas áreas de maior interesse para os pesquisadores tem sido a classificação e a calibração multivariada.

Os problemas de classificação são muito comuns em ciência e engenharia. O reconhecimento de padrões ocorre, quando a partir de um conjunto de treinamento, ou seja, um conjunto para o qual se conhece a categoria a qual se pertence cada amostra, deriva-se regras de classificação, com base em medidas das variáveis relativas de cada espécie [6]. Na classificação, a validação dos modelos é feita através da previsão de categorias para amostras conhecidas.

A calibração multivariada é uma operação que relaciona uma grandeza de saída com uma grandeza de entrada para um sistema em determinadas condições. A calibração multivariada consiste na execução de três passos principais: calibração, validação e previsão.

Para se fazer a calibração multivariada é obtida uma matriz **X**, constituída de inúmeras medidas instrumentais de mesma natureza (como por exemplo, muitos espectros), obtidas para inúmeros padrões de uma ou mais espécies de interesse. Um modelo matemático que melhor correlacione a matriz de resposta **Y** (concentração, por exemplo) a partir dos dados obtidos para a matriz **X** é alcançado.

A validação é o passo seguinte, onde amostras com concentrações conhecidas têm a mesma propriedade prevista para avaliar se o modelo desenvolvido está adequado. Existem dois métodos para a sua condução:

validação interna (quando as próprias amostras de calibração são usadas para a validação) e externa (quando é usado um conjunto distinto, mas com valores Y ainda conhecidos).

Na etapa de previsão, a resposta de interesse para uma amostra desconhecida é obtida utilizando o modelo matemático construído na etapa de calibração e validado posteriormente [5,7].

1.1 Análise de Componentes Principais (PCA)

A análise de componentes principais (PCA) é um tipo de análise exploratória de dados que visa extrair o máximo de informações de uma tabela de dados convertendo-a em gráficos que mostram a relação entre amostras (linhas de uma matriz) e as variáveis (colunas de uma matriz).

O PCA faz uma aproximação da tabela de dados, ou seja, uma matriz **X**, em termos da soma de várias matrizes M_l de posto 1, na qual posto significa um número que expressa a verdadeira dimensionalidade da matriz, como mostra a Figura 1.

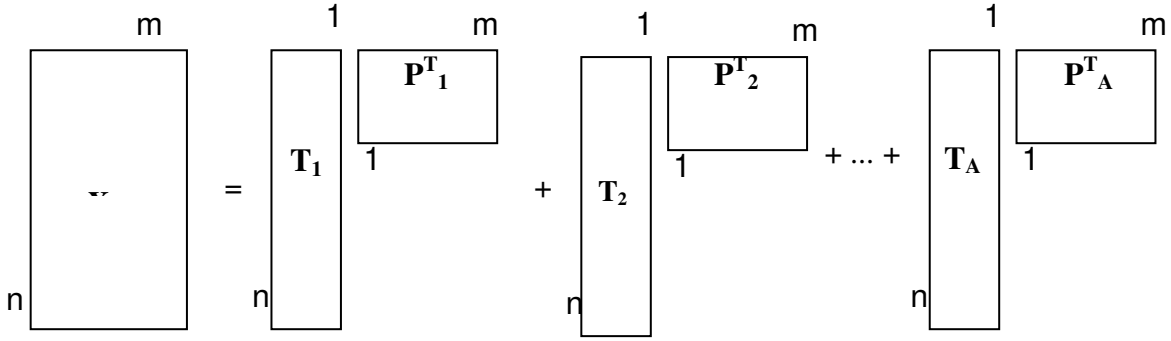


Figura 1. Representação esquemática da decomposição por PCA.

Essa matriz pode ser escrita como produto de dois vetores, escore t_h e peso p_h.

$$\mathbf{X = t_1p_1 + t_2p_2 + ... t_ap_a}, \text{ para “a” componentes principais.} \quad \text{Eq. 1}$$

Outra maneira de escrever a equação anterior é:

$$\mathbf{X = TP^T} \qquad \text{Eq. 2}$$

onde **X** é o conjunto original dos dados com n linhas (amostras) e p colunas (variáveis); **T** é a matriz dos escores com n linhas e d colunas (número de componentes principais, novas coordenadas no novo sistema de eixos) **P** é a matriz dos pesos (informação do peso de cada variável original na formação dos novos eixos) com d linhas e p colunas.

O novo conjunto de variáveis (componentes principais) é a combinação linear das variáveis originais. Os novos eixos são ortogonais entre si e são constituídos em ordem decrescente da quantidade de variância que descrevem. Assim, podemos dizer que o PCA tenta agrupar aquelas variáveis que estão altamente correlacionadas numa nova variável chamada componente principal.

Como usualmente a grande fração da variância é descrita nos primeiros componentes principais, é possível visualizarmos dados pelo gráfico dos escores de um componente contra o outro. Os algoritmos usados para os cálculos com o PCA em química analítica são o NIPALS (do inglês, *Nonlinear Interative Partial Least Squares*) e o SVD (do inglês, *Singular Value Decomposition*) [6].

Os principais objetivos desta técnica são o de encontrar relações entre objetos e classificá-los de acordo com suas similaridades, o que torna possível a detecção de amostras anômalas, ou seja, que não pertencem a nenhuma das categorias conhecidas.

Outro objetivo importante é a redução da dimensão dos dados, que se torna muito útil quando grandes quantidades de informação necessitam ser manipuladas [6,8,9].

1.2 Regressão por Mínimos Quadrados Parciais (PLS)

Esse método de calibração multivariada foi desenvolvido por Herman Wold [6,8,9] na década de 70, baseado em uma relação linear entre as variáveis instrumentais (**X**) e as variáveis de interesse (**Y**). As informações da matriz **X** e da matriz **Y** são usadas ao mesmo tempo na fase de calibração. A matriz dos espectros é decomposta em matrizes de variações dos espectros (*loadings* ou pesos) e a posição das amostras (escores). Os espectros originais podem ser considerados como combinações lineares dos espectros (pesos) onde os escores representam suas contribuições [8].

As matrizes **X** e **Y** podem ser representadas pela Análise de Componentes Principais:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \text{Eq. 3}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \qquad \text{Eq. 4}$$

onde **T** e **U** são as matrizes escores de **X** e **Y**, respectivamente; **P** e **Q** são as matrizes peso de **X** e **Y**; **E** e **F** são os resíduos.

Uma relação linear entre os dois blocos pode ser realizada correlacionando os escores para cada componente de cada vez, utilizando o modelo linear.

$$\mathbf{U_h} = \mathbf{b_h}\mathbf{T_h} \qquad \text{Eq. 5}$$

onde “h” é o numero de componentes principais.

Para que a covariância de **T** e **U** seja maximizada deve-se buscar um modelo onde as matrizes dos resíduos **E** e **F** sejam as menores possíveis e , ao mesmo tempo, conseguir uma relação linear ótima entre **t** e **u**.

No PLS isto é obtido por uma leve mudança nos valores dos escores, de forma a produzir a melhor relação possível. Nesta etapa as componentes principais deixam de possuir este nome sendo chamadas então de variáveis latentes, pois elas não mais descrevem a máxima variância dos dados.

O algoritmo utilizado para a estimativa dos vetores coeficientes de determinação b é o NIPALS ou SIMPLS (do inglês, *Straight Foward*

Implementation of Statistically Inspired Modification of PLS). Chamando A_{max} do número máximo de fatores a serem computados pelo algoritmo, executa-se os seguintes passos para cada um dos fatores a=1,2...A_{max}.

1-encontrar vetor peso $\mathbf{w_a}$ para maximização da covariância entre a combinação linear $\mathbf{x_{a-1}w_a}$ e **y**, com condição que $\mathbf{w_a^T w_a}$=1. Isto corresponde encontrar o vetor unitário $\mathbf{w_a}$ que maximiza $\mathbf{w_a^T x_{a-1}^T y_{a-1}}$, ou seja, a variância escalada entre $\mathbf{x_{a-1}}$ e $\mathbf{y_{a-1}}$.

2-encontrar os escore, $\mathbf{t_a}$ como projeção de $\mathbf{X_{a-1}}$ em $\mathbf{w_a}$, isto é,

$$\mathbf{t_a = X_{a-1} W_a};$$ Eq. 6

3-realizar a regressão de $\mathbf{X_{a-1}}$ em $\mathbf{t_a}$ para encontrar os vetores pesos de **X**

$$\mathbf{p_a^T = x_{a-1}^T t_a / t_a^T t_a}$$ Eq. 7

4-realizar regressão de $\mathbf{y_{a-1}}$ em $\mathbf{t_a}$ para encontrar vetores pesos de **Y**

$$\mathbf{q_a = y_{a-1} t_a / t_a^T t_a;}$$ Eq. 8

5-subtrair $\mathbf{t_a p_a^T}$ de $\mathbf{X_{a-1}}$ e chamar esta nova matriz de $\mathbf{X_a}$ e subtrair $\mathbf{t_a q_a}$ de $\mathbf{y_{a-1}}$ e chamar esta nova matriz de $\mathbf{y_a}$;

6-otimizar o número de fatores por validação cruzada, e considerar as seguintes matrizes

W={wa};

P={pa};

Para **a** número de fatores otimizados;

7-calcula-se, então, os coeficientes de determinação **b** através da seguinte relação:

$$\mathbf{b = W(P^T W)^{-1} q}$$ Eq. 9

em que **W** é a matriz de pesos do PLS e **q** os *loadings* de **Y**.

Após o modelo ter sido desenvolvido é feita a validação do modelo com novas amostras. A validação cruzada é baseada na avaliação da grandeza dos erros de previsão comparando com os valores das variáveis dependentes das amostras do conjunto de calibração com as respectivas previsões, quando as mesmas não participam na construção do modelo de regressão. Na validação cruzada “um por vez” o PLS separa uma das amostras de calibração e a usa como

validação. Isto é feito até que todas as amostras de calibração sejam usadas como amostra de validação. Em paralelo é realizada a validação com um conjunto externo que deve apresentar amostras com valores que compreendam o intervalo de dados do conjunto de calibração e que devem apresentar performance muito semelhante em todos os parâmetros de avaliação dos modelos de calibração.

1.3 *Soft Independent Modeling of Class Analogy* (SIMCA)

Nesse método de classificação, são desenvolvidos modelos baseados na Análise de Componentes Principais (PCA) para cada classe previamente conhecida. Em termos geométricos, cada modelo descreve um envelope ou "caixa" ao redor de cada classe de modo que objetos desconhecidos (novas amostras) podem ser classificados como pertencentes àquela classe em particular se ficarem dentro desses envelopes. A dimensão de cada envelope é dada pela variância das amostras em torno dos componentes principais [10,11].

O SIMCA é um método para classificação que considera informações da distribuição da população, estima um grau de confiança da classificação e pode prever novas amostras como pertencentes a uma ou mais classes ou nenhuma classe. Para fazer a classificação o SIMCA utiliza o espaço das componentes principais de cada classe. Desta forma, a classe *n* passa a ser representada pela equação a seguir [12]:

$$\mathbf{X_n} = \mathbf{T_n P_n} + \mathbf{E} \qquad \text{Eq. 10}$$

onde, $\mathbf{X}_n$ são os dados da classe, $\mathbf{T}_n$ a matriz contendo as coordenadas nas componentes principais da classe *n* (matriz de escores), $\mathbf{P}_n$ a matriz de transformação linear (matriz de *loadings*) e $\mathbf{E}_n$ a matriz de resíduos.

Na construção do modelo de classificação o SIMCA calcula, para cada classe em separado, o desvio padrão dos resíduos. Para o espaço descrito pelas componentes principais, são calculadas as variâncias das amostras, em cada eixo. Estes dois parâmetros são usados na classificação de novas amostras. O

objetivo do SIMCA é criar um espaço limitado para cada classe. Isto pode ser mais bem compreendido para uma classe descrita por duas componentes principais. Em termos geométricos, os resíduos desta classe correspondem às distâncias das amostras ao plano das componentes principais. Desta forma, o cálculo do desvio padrão dos resíduos dá origem a dois planos paralelos ao destes componentes, isto é, um acima e outro abaixo. Considerando a variância em cada componente principal e os planos, referentes ao desvio padrão dos resíduos, pode-se dizer que a classe está limitada por uma caixa, uma hipercaixa, no caso de três ou mais componentes, e um cilindro para uma componente principal [12,11].

A classificação de uma nova amostra é feita através de sua projeção nas componentes principais de cada classe, onde são calculados as variâncias e seu resíduo. Assim, naquelas classes onde o resíduo é menor ou igual, o mesmo é válido para as variâncias, a amostra é classificada positivamente. Com isto, a amostra pode ser colocada em uma ou mais classes. Em caso contrário, desvio ou variâncias maiores, a amostra é classificada como não pertencente à classe [12]. A Figura 2 apresenta um exemplo de SIMCA com amostras pertencentes a três conjuntos distintos.

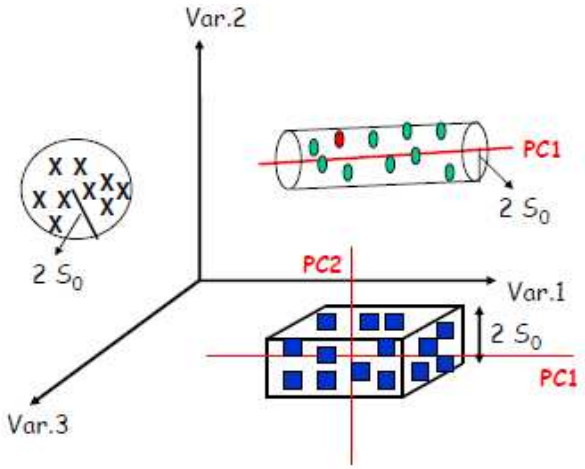


Figura 2. Representação gráfica de um modelo SIMCA

1.4 Pré-processamento dos sinais analíticos

Uma etapa importante no desenvolvimento de um modelo de calibração é a etapa de pré-processamento. Muitas vezes os dados a serem modelados são expressos em grandezas diferentes, apresentam muitos ruídos ou variações de linha base que podem prejudicar o desempenho do modelo. Assim tratamentos são realizados nos dados antes do desenvolvimento do modelo de calibração.

1.4.1 Aplicação de derivada

A aplicação da primeira ou segunda derivada sobre os dados espectrais brutos é um procedimento que pode destacar ombros espectrais, bem como minimizar o efeito de inclinações provocadas na linha de base dos espectros, devido à morfologia das partículas [9,13].

Ao aplicarmos as operações de derivação aos espectros, as informações contidas ao longo dos diferentes comprimentos de onda são geralmente acentuadas. Não só os sinais espectrais, mas também os ruídos tornam-se acentuados, portanto, deve-se ter cuidado com a qualidade dos espectros com os quais se deseja aplicar o cálculo das derivadas [13].

1.4.2 Dados centrados na média

A centralização na média [14] consiste em fazer com que para cada variável seus valores tenham média zero. Para centrar os dados na média, obtêm-se para cada coluna o valor médio e, em seguida, subtrai-se este valor de cada variável dessa mesma coluna. Desta forma, ocorre a mudança do sistema de coordenadas para o centro dos dados. A Equação 11 é utilizada para centrar os dados na média.

$$x_{(i,j\ cm)} = x_{i,j} - \overline{x_j} \qquad \text{Eq. 11}$$

em que, $x_{(i,j\ cm)}$, corresponde ao valor centrado na média para a variável *j* na amostra *i* ; $x_{i,j}$, é o valor da variável *j* na amostra *i* e $\overline{x_j}$ é a média das amostras na coluna calculada pela Equação 12.

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \qquad \text{Eq. 12}$$

onde *n* representa o número de amostras.

1.4.3 Correção do espalhamento multiplicativo (MSC)

O método de correção de espalhamento multiplicativo (MSC - do inglês, M*ultiple Scatering Correction*) [15] é comumente aplicado em espectroscopia para a correção de linha base, proveniente principalmente da não homogeneidade da distribuição de partículas na matriz.

Este método assume que os comprimentos de onda da luz espalhada possuem uma dependência distinta entre a luz espalhada e a absorvida pelos constituintes da amostra. Portanto teoricamente, é possível separar estes dois sinais. Este método tenta remover o efeito do espalhamento pela linearização de cada espectro por um espectro ideal. Para efeito de cálculo, considera-se que o espectro ideal é o espectro médio do conjunto de dados para o qual se deseja realizar a correção da linha base. Em seguida, utiliza-se uma regressão linear para calcular o coeficiente angular e linear do gráfico entre o espectro ideal e o espectro que vai ser corrigido. O espectro corrigido é calculado subtraindo cada ponto do espectro pelo valor do coeficiente linear e dividindo este valor pelo coeficiente angular [16].

Matematicamente, e resumindo, a correção é feita da seguinte forma:

1. A partir do conjunto total de espectros, calcula-se o espectro médio $\overline{x_i}$;

2. Faz-se a regressão linear para cada um dos *k* espectros (x_{ik}) do conjunto total de espectros, contra o espectro médio, sobre todos os *i* comprimentos de onda:

$$x_{ik} = v_k\overline{x_i} + u_k \qquad \text{Eq. 13}$$

Onde v_k é o coeficiente angular e u_k o coeficiente linear.

3. Correção final:

$$x_{ik}^{(corrigido)} = \frac{\left(x_{ik}^{(\text{não } corrigido)} - u_k\right)}{v_k} \qquad \text{Eq. 14}$$

1.4.4 Transformação padrão normal de variação (SNV)

Espectros na região do infravermelho podem apresentar problemas de linha base devido principalmente ao espalhamento de luz. O espalhamento é fortemente dependente do comprimento de onda da luz, do tamanho das partículas, do índice de refração etc. Para minimizar este efeito, é necessário o uso de técnicas como a transformação padrão de variação (SNV – do inglês *Standard Normal Variate*) [17]. Esta técnica é aplicada para corrigir os efeitos do espalhamento multiplicativo e o tamanho da partícula, de maneira análoga à correção de espalhamento multiplicativo (MSC) [18].

Apesar do MSC e SNV terem a mesma finalidade, ou seja, corrigir a linha base espectral, estas duas técnicas são bem diferentes. O SNV não necessita de um espectro ideal, ou seja, de um espectro médio para fazer a correção dos espectros. A correção é realizada pela normalização de cada espectro para o seu próprio desvio padrão p, conforme ilustrado pelas equações 15 e 16 a seguir:
Média do espectro:

$$\bar{x} = \sum_{j=1}^{p} \frac{X_i}{p} \qquad \text{Eq. 15}$$

Espectro corrigido

$$x_{i\,(SNV)} = \frac{(X_i - \overline{X_i})}{\sqrt{\frac{\sum_{j=1}^{p}(X_i - \overline{X_i})^2}{p-1}}}$$ Eq. 16

em que **X** representa uma matriz com *n* espectros, *p* representa o número de pontos no espectro, x_i é a média do vetor contendo o espectro e $x_{i\,(SNV)}$ é o espectro corrigido.

1.4.5 Normalização dos espectros

A normalização [19] é usada principalmente para remover variação sistemática, geralmente associada com a espessura da amostra. Na normalização, dividem-se cada uma das variáveis de uma dada amostra *i* por um fator de normalização, ou seja, pela norma da amostra *i*, representada por x_i. O resultado é que todas as amostras estarão numa mesma escala.

$$x_{ij\,(norm)} = \frac{X_{ij}}{\|X_i\|} \qquad j=1,2...J. \qquad \text{Eq. 17}$$

As normas utilizadas são:

$$\|X_i\|_\infty = max_{1\le j\le J}\left|X_{ij}\right| \qquad \text{norma sup} \qquad \text{Eq. 18}$$

$$\|X_i\|_1 = \sum_{j=1}^{J}\left\|X_{ij}\right\| \qquad \text{norma } l_1 \qquad \text{Eq. 19}$$

$$\|X_i\|_2 = \sum_{j=1}^{J}\left\|X_{ij}^2\right\| \qquad \text{norma } l_2 \qquad \text{Eq. 20}$$

- Normalização pela norma sup: a resposta máxima de cada uma das amostras se torna igual a 1.
- Normalização pela norma l_1: a área sob cada um dos espectros é unitária.

- Normalização pela norma l_2: cada espectro terá comprimento igual a 1.

1.4.6 Correção Ortogonal de Sinal (OCS)

Para remover variações sistemáticas indesejáveis em dados analíticos, dois tipos de pré-processamentos são os mais comumente encontrados na literatura, a diferenciação e a correção de sinal.

Essas correções de sinal são diferentes casos de filtragem, onde um sinal passa por um filtro, que é uma função matemática, para ter suas características melhoradas. Entretanto nem sempre é fácil construir estes filtros e os objetivos da filtragem são muitas vezes bastante vagos.

Mesmo no caso da calibração multivariada, onde o objetivo pode ser especificado em termos de erros de previsão mais baixos, é difícil construir filtros que, de fato, melhorem estas propriedades dos dados [20].

A correção ortogonal de sinal (OSC, do inglês *Orthogonal Signal Correction*) [20,21] parte da ideia de retirar da matriz espectral **X** apenas a parte que está definitivamente não relacionada a uma característica **Y**. Isto é feito garantindo que a parte removida é matematicamente ortogonal a **Y**.

Baseado no algoritmo NIPALS, o OSC remove variações químicas e de fundo, deixando o espectro mais simples para o desenvolvimento do modelo de calibração. O algoritmo básico do OSC é:

1. Componentes principais são calculados de acordo com o NIPALS:

$$\mathbf{X=tp^T + E} \qquad \text{Eq. 21}$$

2. O escore é ortogonalizado contra as variáveis a serem previstas (**y**), resultando em $\mathbf{t^*}$

$$\mathbf{t^*=(I-Y)((Y^TY)^{-1}Y^T)t} \qquad \text{Eq. 22}$$

3. Após a ortogonalização, os pesos **w** do PLS são calculados de maneira que:

$$\mathbf{t^*=Xw} \qquad \text{Eq. 23}$$

4. Os pesos **w** são utilizados para minimizar a covariância entre **X** e **y.**
5. Dados espectrais ortogonais:

$$\mathbf{X = t^{*} p^{*} + E} \qquad \text{Eq. 24}$$

6. Os resíduos **E** constituem os dados filtrados após a remoção dos componentes ortogonais a **Y**:

$$\mathbf{E = X - t^{*} p^{*}} \qquad \text{Eq. 25}$$

$$\mathbf{X_{osc} = E} \qquad \text{Eq. 26}$$

1.5 Algoritmo genético

Algoritmo genético (GA) [22], é uma metodologia de busca de otimização baseada numa analogia direta à teoria da seleção natural e genética em sistemas biológicos de Darwin. O GA trabalha com um grupo de soluções candidatas chamado de População. Baseada no principio darwiniano de "sobrevivência do mais adaptado", o GA obtém a solução ótima depois de uma série de cálculos iterativos.

O GA básico envolve cinco passos: codificação das variáveis, criação da população inicial, avaliação da resposta, cruzamento e mutação. A implementação do GA na seleção de variáveis difere-se das aplicações normalmente realizadas no que tange à codificação do problema e a função de resposta, já que as outras etapas permanecem inalteradas. No caso da seleção de variáveis, considera-se que o cromossomo possui "p" genes, onde cada gene representa uma das variáveis do sinal analítico (espectro, por exemplo) sendo então o número de genes igual ao número de variáveis contidas nesse sinal.

Na seleção de variáveis utiliza-se o auxílio do código binário (0,1) para codificar o problema. Cada gene pode assumir o valor um ou zero. Quando a posição referente a uma determinada variável for igual a um, implicará na seleção desta variável, se a posição contiver o valor zero, a variável não será selecionada.

A partir da população inicial, uma nova população a qual pode ser considerada como próxima geração, é obtida pelo cruzamento aleatório entre material genético de cromossomos diferentes. No cruzamento, dois cromossomos pais são divididos geralmente em duas ou três partes, cada uma escolhida aleatoriamente, que são cruzadas e combinadas para formar dois cromossomos

filhos que substituirão os cromossomos pai dentro de uma nova geração (Figura 3). Uma nova avaliação é realizada e os cromossomos com valores de aptidão maiores têm uma probabilidade de reprodução maior que os cromossomos com valores menores, tudo para melhorar a aptidão global da população [22-24].

Mutações podem ser incorporadas ao modelo e são, às vezes, necessárias para superar alguns problemas na população, sendo utilizadas para:

- dar nova informação genética à população, ou seja, uma variável não selecionada em quaisquer dos cromossomos originais, nunca seria selecionada na próxima geração se mutações não tivessem presentes;

- prevenir que a população se sature com cromossomos semelhantes (convergência prematura).

Uma mutação nada mais é que a inversão de um gene no cromossomo. Ainda usando o exemplo para dois cromossomos pai com seis genes (variáveis) podemos representar a mutação do gene 4 como visto na Figura 3.

O algoritmo é repetido até que a condição de término é cumprida. A condição de término é baseada no critério de convergência, em que o algoritmo é encerrado quando uma certa porcentagem dos cromossomos for idêntica ou quando um determinado número de gerações é atingido [22,24].

Na avaliação da resposta, ou seja, aptidão deve-se encontrar o valor associado à eficiência de cada cromossomo relacionado ao sistema de interesse, sendo o resultado mais importante no procedimento do algoritmo genético. A aptidão é uma característica intrínseca ao indivíduo, que representa sua habilidade de produzir a melhor resposta. O objetivo é encontrar o menor erro possível, e este será o responsável direto pela vida ou morte dos indivíduos [23]. Todo ciclo evolucionário é mostrado na Figura 4.

Como vantagens, devemos salientar a capacidade deste algoritmo em lidar com grandes espaços de busca e obter a melhor solução local em relação a outros algoritmos.

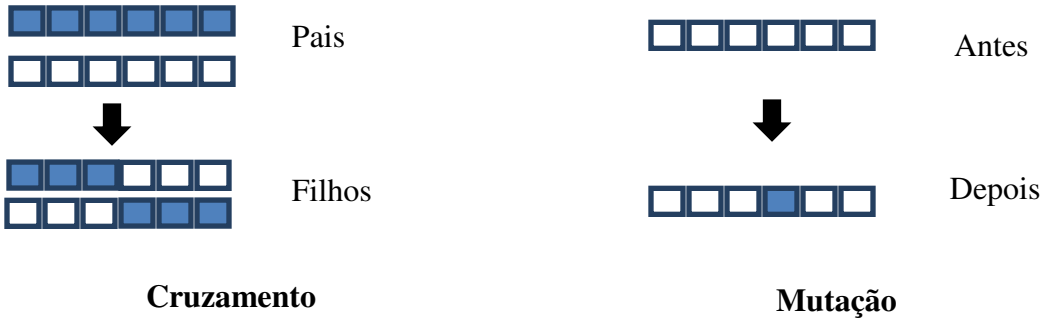


Figura 3. Operações genéticas

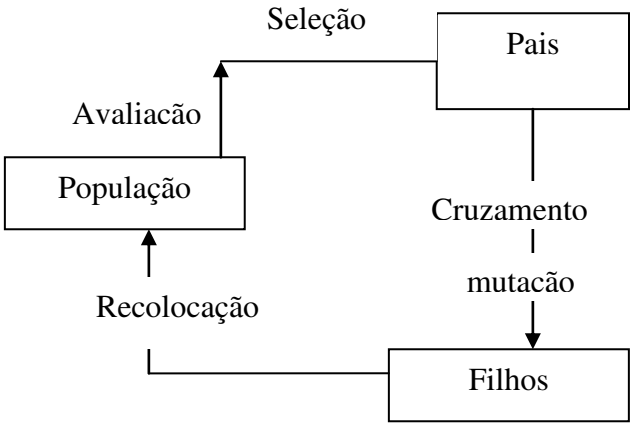


Figura 4. Ciclo Evolucionário

1.6 Raiz quadrada do erro médio quadrático

Comumente em aplicações com Calibração Multivariada utiliza-se o parâmetro RMSEP (*Root Mean Square Error of Prediction*) ou Raiz Quadrada do Erro Médio Quadrático de Previsão, que expressa o grau de concordância entre os valores estimados por um modelo previamente construído e o valor considerado real ou de referência.

$$RMSE = \sqrt{\frac{\sum (y_p - y_r)^2}{n}} \qquad \text{Eq. 27}$$

onde *yp* são os valores previstos pelo modelo, *yr* são os valores de referência e *n* é o número de amostras utilizadas no conjunto de validação.

O RMSEC (*Root Mean Square Error of Calibration*) ou raiz quadrada do erro médio quadrático de calibração nos fornece informação sobre o ajuste do modelo aos dados de calibração. O RMSEC é calculado como na equação 27, sendo utilizados os *n* valores previstos no conjunto de calibração.

O RMSECV (*Root Mean Square Error of Cross-Validation*) ou raiz quadrada do erro médio quadrático da validação cruzada fornece uma medida sobre a habilidade do modelo em prever novas amostras. O RMSECV é definido como na equação 27, com a exceção de que yp são as previsões para amostras não incluídas no modelo.

Como pode ser observado na equação 27, o RMSEP é uma medida de dispersão semelhante ao desvio padrão, mas que mede a dispersão entre os valores estimados pelo modelo e de referência. Outra propriedade que se assemelha à do desvio padrão é que o RMSEP é uma medida que considera apenas erros aleatórios, que é uma decorrência da elevação dos erros ao quadrado. Por exemplo, considerando os resultados de dois métodos distintos, supondo que um apresente erros sistemáticos negativos e o outro tenha erros com o mesmo valor em módulo mas que sejam distribuídos de forma aleatória, ambos fornecem os mesmos valores de RMSEP. Assim, a constatação de que dois RMSEP são estatisticamente equivalentes por meio de um teste-F torna possível afirmar que os erros médios na estimativa da propriedade de interesse dos dois métodos são equivalentes não podendo ser utilizada para inferir sobre a exatidão do método [25-27].

1.7 Comparação de conjuntos de dados – Teste F

Em trabalhos experimentais, especialmente no desenvolvimento de um novo procedimento de análise, é comum realizar-se uma avaliação estatística dos resultados obtidos, visando identificar a existência de uma diferença significativa

na variância entre este conjunto de respostas e outro conjunto obtido por um procedimento de referência. Esta avaliação é feita usando-se o teste F. Este teste usa a razão das variâncias ao quadrado S_1^2 é e S_2^2 dos dois conjuntos de respostas para estabelecer se efetivamente existe uma diferença estatisticamente significativa entre os valores que estão sendo comparados. Nos casos apresentados nesta tese as variâncias são os valores RMSE obtidos, sendo o S_1 sempre o RMSEP de maior valor. O valor de F é calculado pela seguinte expressão:

$$F = \frac{S_1^2}{S_2^2} \qquad \text{Eq. 28}$$

O valor de F obtido é comparado a valores críticos calculados para um determinado nível de confiança. Quando o valor experimental de F excede o valor crítico tabelado, então a diferença na variância é tomada como estatisticamente significante [25-27].

Capítulo II

2. Máquinas de vetor de suporte (SVM)

O algoritmo Máquinas de Vetor de Suporte (SVM, do inglês *Support Vector Machines*) [28] pode ser usado para classificação de padrões e calibração e foi introduzido primeiramente nas áreas de engenharia. Na área da quimiometria as aplicações encontradas na literatura ainda são escassas.

A ideia principal de uma máquina de vetor de suporte é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima.

A Máquina de Vetor de Suporte é uma implementação do método de minimização estrutural de risco. Este princípio é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre dados de teste (isto é, a taxa de generalização) é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis (V-C); no caso de padrões separáveis, o SVM produz um valor de zero para o primeiro termo e minimiza o segundo. Consequentemente, os SVM podem fornecer um bom desempenho de generalização em problemas de classificação de padrões, apesar do fato de que ela não incorpora conhecimento do domínio do problema.

Uma noção que é central à construção do algoritmo de SVM é o núcleo interno entre um "vetor de suporte" $\mathbf{x_i}$ e o vetor $\mathbf{x}$ retirado do espaço de entrada. Os vetores de suporte consistem de um pequeno subconjunto dos dados de treinamento extraído pelo algoritmo [29].

Modelos matemáticos com capacidade de aproximação universal, como as redes neurais artificiais, ainda não são dotadas de algoritmos de treinamento capazes de maximizar a capacidade de generalização de uma forma sistemática, o que pode levar a um sobreajuste do modelo aos dados. Como não são conhecidas as não-linearidades presentes e a complexidade intrínseca do problema, os algoritmos de otimização e as ferramentas estatísticas utilizadas para seleção de modelos podem induzir modelos com baixa capacidade de generalização, assim o SVM se torna uma poderosa alternativa para resolver problemas de classificação e calibração.

Sendo assim, as principais vantagens do SVM em suas aplicações são:
-Elevada capacidade de generalização, evitando o sobreajuste;
-Robustez em grandes dimensões, possibilitando aplicação de SVMs em vetores de características de grandes dimensões;
-Convexidade da função objetivo; a aplicação das SVMs implica na otimização de uma função quadrática, que possui apenas um mínimo (Figura 5);
-Teoria bem estabelecida dentro da Matemática e Estatística [28].

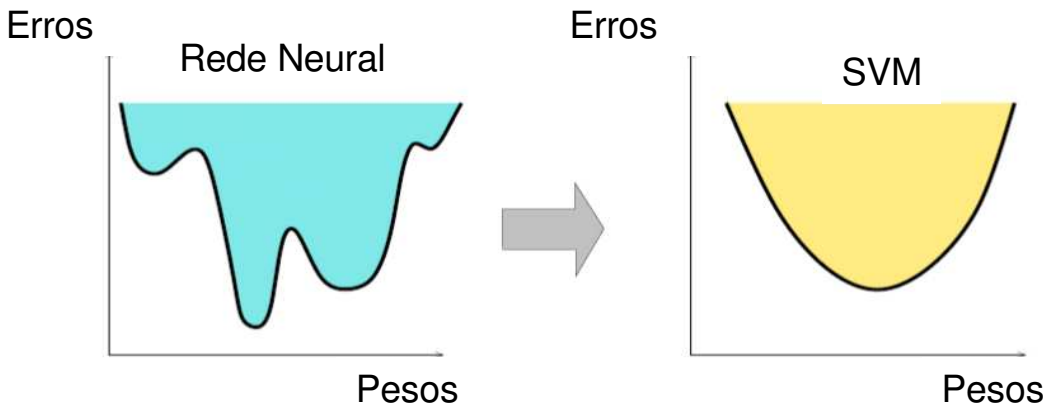


Figura 5. Convexidade da função objetivo.

2.1 Hiperplano ótimo

Considerando uma amostra de treinamento $\mathbf{x}_i$, assumimos que os subconjuntos representados por d_i=+1 e d_i=-1 são "linearmente separáveis" (Figura 6A). A equação de uma superfície de decisão na forma de um hiperplano que realiza esta separação é:

$$\left(\mathbf{w}^T\mathbf{x}+b\right)=0 \qquad \text{Eq. 29}$$

Onde **x** é um vetor de entrada, **w** é um vetor peso ajustável e b é um bias. Podemos assim escrever:

$$\mathbf{w}^T \mathbf{x}_i + b \geq 0, d_1 = 1 \quad \text{Eq. 30}$$

$$\mathbf{w}^T \mathbf{x}_i + b < 0, d_1 = -1 \quad \text{Eq. 31}$$

Para um dado vetor de peso **w** e bias b, a separação entre o hiperplano definido na Eq 29 e o ponto de dado mais próximo é denominada a margem de separação, representada por ρ. O objetivo de uma SVM é encontrar o hiperplano particular para qual a margem de separação é máxima. Sob esta condição, a superfície de decisão é referida como o hiperplano ótimo.

Considere que $\mathbf{w}_0$ e b_0 representem os valores ótimos do vetor peso e do bias, o hiperplano ótimo, representando uma superfície de decisão linear multidimensional no espaço de entrada, é definido por:

$$\boldsymbol{w}_0^T \boldsymbol{x} + b_0 = 0 \quad \text{Eq. 32}$$

A função discriminante:

$$g(x) = \boldsymbol{w}_0^T \boldsymbol{x} + b_0 \quad \text{Eq. 33}$$

fornece uma medida algébrica da distancia de x até o hiperplano. Dado um conjunto de treinamento a questão a resolver é encontrar os parâmetros ótimos $\mathbf{w}_0$ e b_0 para o hiperplano ótimo, onde o par satisfaça a restrição:

$$\boldsymbol{w}_0^T \boldsymbol{x_i} + b_0 \geq 1 \text{ para } d_i=+1 \quad \text{Eq. 34}$$

$$\boldsymbol{w}_0^T \boldsymbol{x_i} + b_0 \leq 1 \text{ para } d_i=-1 \quad \text{Eq. 35}$$

Os pontos de dados particulares (x_i,d_i) para as quais a Equação 34 ou a Equação 35 é satisfeita com sinal de igualdade são chamados de vetores de suporte. Em termos conceituais, os vetores de suporte são aqueles pontos de

dados que se encontram mais próximos da superfície de decisão e são, portanto, os mais difíceis de classificar (Figura 6B).

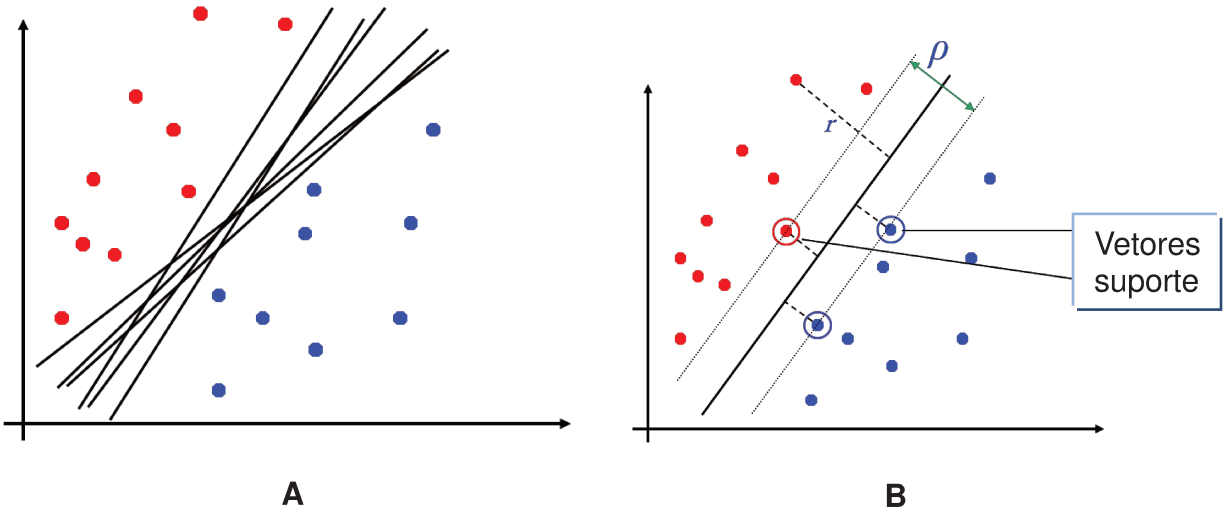


Figura 6. A) Subconjuntos linearmente separáveis, **B)** Vetores de suporte.

Considerando a distância algébrica de um vetor de suporte x(s) ao hiperplano ótimo

$$\boldsymbol{r} = \frac{g(\boldsymbol{x}(s))}{\|\boldsymbol{w}_0\|} = \frac{1}{\|\boldsymbol{w}_0\|} \text{ se d(s) = +1 e} \qquad \text{Eq. 36}$$

$$\boldsymbol{r} = -\frac{1}{\|\boldsymbol{w}_0\|} \text{ se d(s)=-1} \qquad \text{Eq. 37}$$

onde o sinal positivo indica que x(s) se encontra no lado positivo do hiperplano ótimo e o sinal negativo indica que x(s) está do lado negativo do hiperplano ótimo. Considere que ρ represente o valor ótimo da margem da separação entre duas classes que constituem o conjunto de treinamento τ. Então, das Equações 36 e 37 resulta que:

$$\rho = 2r \qquad \text{Eq. 38}$$

$$\rho = \frac{2}{\|w_0\|} \qquad \text{Eq. 39}$$

A Equação 39 afirma que maximizar a margem de separação entre classes é equivalente a minimizar a norma euclidiana do vetor peso **w**. Em resumo, o hiperplano ótimo definido é único no sentido de que o vetor peso $\mathbf{w}_0$ fornece a máxima separação entre exemplos positivos e negativos. Esta condição ótima é alcançada minimizando-se a norma euclidiana do vetor peso **w**.

O objetivo do SVM passa a ser encontrar um hiperplano ótimo para um conjunto de treinamento. O problema de otimização restrito que temos que resolver pode ser formulado como:

$$d_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 \text{ para i=1,2,3} \qquad \text{Eq. 40}$$

de maneira que o vetor peso seja minimizado:

$$\varphi(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \qquad \text{Eq. 41}$$

Essa função custo (φ) é uma função convexa de $\boldsymbol{w}$ e, portanto, pode-se encontrar sempre uma solução para o seu mínimo [29].

2.2 SVM para dados não separáveis

A discussão até agora enfocou padrões linearmente separáveis. Dado um conjunto de dados de treinamento não separáveis não é possível construir um hiperplano de separação sem nos defrontarmos com erros de classificação. Apesar disso, é possível encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação, calculado sobre o conjunto de treinamento.

Diz-se que a margem de separação entre classes é suave se uma amostra violar a seguinte condição [29]:

$$\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right)y_i \geq 1 \qquad \text{Eq. 42}$$

Neste ponto, é introduzido um novo conjunto de variáveis escalares não negativas, ξ_i, que são chamadas de variáveis "soltas" e determinam a superfície de decisão (Figura 7); elas medem o desvio de uma amostra da condição ideal de separabilidade de padrões [29,30].

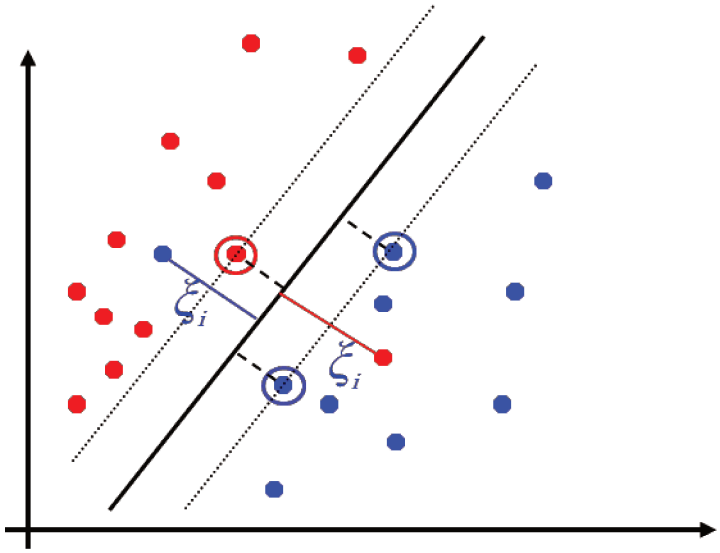


Figura 7. Variáveis "soltas"

Para $0 \leq \xi_i \leq 1$, o dado encontra-se dentro da região de separação, mas no lado correto da superfície de decisão. Para $\xi_i > 1$, ele se encontra no lado errado do hiperplano de separação. Os vetores de suporte são, portanto, aqueles pontos de dados particulares que satisfazem a equação:

$$d_i\left(w^T x_i + b\right) \geq 1 - \xi_i \quad \text{i = 1,2,...,N} \qquad \text{Eq. 43}$$

Mesmo se $\xi_i > 0$, a superfície de decisão não será alterada. Assim, os vetores de suporte são definidos exatamente do mesmo modo, tanto para o caso de serem linearmente separáveis, como os que não os são [29,30].

O objetivo passa a ser encontrar um hiperplano de separação para qual o erro de classificação do próprio conjunto de treinamento é o mínimo possível. Isso pode ser feito minimizando a função abaixo em relação ao peso **w**:

$$\phi(\mathbf{w},\xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\left(\sum_{i=1}^{N}\xi_i\right) \qquad \text{Eq. 44}$$

O parâmetro γ controla o compromisso entre a complexidade da máquina e o número de pontos não-separáveis; por isso, pode ser visto como uma forma de parâmetro de "regularização". Esse parâmetro deve ser otimizado pelo usuário.

Tem-se agora o problema de encontrar **w** de tal maneira que se respeite a restrição da equação 43 e $\xi \geq 0$. Fazendo isso, a norma quadrada de **w** é tratada como uma quantidade a ser minimizada simultaneamente aos dados não separáveis, e não como uma restrição imposta sobre a minimização do número de pontos não separáveis. O problema de otimização para padrões não-separáveis assim formulado inclui o problema para padrões linearmente separáveis como um caso especial [29].

Esse problema de otimização restrita pode ser resolvido pelo método dos multiplicadores de Lagrange [31].

$$L(\mathbf{w},b,\xi,\alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \gamma\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left\{\mathbf{w}^T\mathbf{x}_i + b + \xi_i - y_i\right\} \quad \text{Eq. 45}$$

Em que:

$$y_i = \begin{bmatrix} y \\ y \\ \vdots \\ y_N \end{bmatrix},\ e_{i=}\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_N \end{bmatrix} \text{ e } \alpha_i = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix}. \qquad \text{Eq. 46}$$

A função custo definida antes se encontra nos dois primeiros termos desta função de Lagrange (L) (Eq 45), mas essa equação possui multiplicadores de restrição, os chamados multiplicadores de Lagrange (α_i), sendo cada um correspondente a um dado de treinamento. Para obter a solução final as primeiras derivadas parciais são acertadas para zero e combinações lineares dos dados de treinamento são obtidas. Os coeficientes de determinação (**w**) podem ser escritos

como uma expansão dos multiplicadores de Lagrange com os respectivos dados de treinamento (x_i):

$$\frac{\partial L(w,b,\xi,\alpha)}{\partial w} = w - \sum_{i=1}^{N} \alpha_i \phi(x_i) = 0 \therefore w = \sum_{i=1}^{N} \alpha_i \phi(x_i) \quad \text{Eq. 47}$$

$$\frac{\partial L(w,b,\xi,\alpha)}{\partial \xi} = \xi - \alpha = 0 \therefore \alpha = \gamma \xi \quad \text{Eq. 48}$$

Então a solução ótima para o vetor peso é dada por:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \phi(x_i) \quad \text{Eq. 49}$$

onde *N* é o número de vetores de suporte. Um importante resultado desta aproximação é que os pesos (**w**) podem ser escritos como combinações lineares dos multiplicadores de Lagrange com os dados de treinamento correspondentes ($\mathbf{x}_i$). Então, colocando essa expressão na reta de regressão original **(y**=**wx**+b), o seguinte resultado é obtido:

$$y = \sum_{i=1}^{N} \alpha_i \, \phi(x_i)^T \, \phi(x) + b = \sum_{i=1}^{N} \alpha_i \left\langle \phi(x_i)^T, \phi(x) \right\rangle + b \quad \text{Eq. 50}$$

onde o produto interno de x_i e x é indicado por $\left\langle \phi(x_i)^T, \phi(x) \right\rangle$ [31].

2.3 SVM para dados não lineares

O modelo final pode ser descrito como uma combinação linear dos produtos internos entre os dados de treinamento e do novo dado (**x**). Isso é importante por duas razões, sendo elas: a dimensão dos dados não aparece no problema e é mais fácil permitir regressão não-linear como uma extensão da aproximação linear [32].

Por fim o produto interno $\langle \phi(x_i)^T, \phi(x) \rangle$ é substituído por uma função de Kernel K(xi,x). A função Kernel representa a relação entre o dado de entrada e a propriedade de saída a ser modelada [33]. Essa função determina tanto o mapeamento não-linear, x$\rightarrow$ ϕ (x), quanto o produto interno correspondente $\phi(x_i)^T \phi(x)$. Isso leva à seguinte função de regressão não-linear:

$$y = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b \qquad \text{Eq. 51}$$

Cada kernel é associado com um parâmetro específico, para as funções polinomiais e de função radial de base (RBF) os parâmetros são o grau do polinômio (d) e a largura da função Gaussiana (σ), respectivamente. Assim, ao invés de calcular um mapeamento específico para cada dimensão dos dados, uma função Kernel apropriada é selecionada e seu parâmetro específico é otimizado [32].

A função Kernel transforma o espaço de entrada em um espaço de características de alta dimensão onde a solução do problema pode ser representada como sendo um problema linear, como mostra a Figura 8 [33].

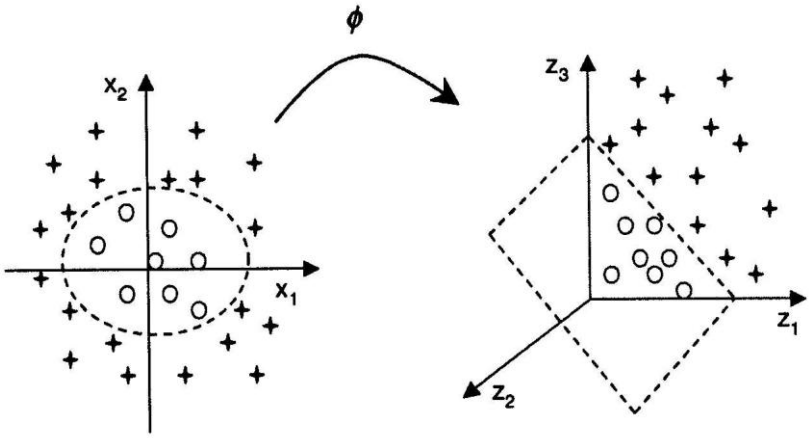


Figura 8. Mudança do espaço dos dados pela função Kernel (ϕ).

A função Kernel mais comumente utilizada é a função radial de base (RBF) [29] :

$$K = \exp\left(- \left\| x_i - x \right\|^2 \Big/ 2\sigma^2 \right) \qquad \text{Eq. 52}$$

Enquanto o produto interno Kernel é uma medida de similaridade entre dois vetores, o Kernel RBF é conceitualmente uma medida não-linear de similaridade. O ajuste do parâmetro de variância (σ^2) muda a largura da gaussiana e o grau da não-linearidade que pode ser modelada. Conforme σ^2 é aumentado o Kernel força o modelo para uma solução linear [34]. A dimensionalidade do espaço (oculto) de características é feito propositadamente muito grande para a construção de uma superfície de decisão na forma de um hiperplano naquele espaço. Os parâmetros γ e σ^2 devem ser sintonizados simultaneamente.

A máquina de vetor de suporte fornece um método para controlar a complexidade do modelo independentemente da dimensionalidade. Em particular, o problema da complexidade do modelo é resolvido em um espaço de alta dimensionalidade usando um hiperplano penalizado definido no espaço (oculto) de características como a superfície de decisão; o resultado é um bom desempenho de decisão.

Mais importante que isso, usando um núcleo de um produto interno adequado, uma SVM calcula automaticamente todos os parâmetros importantes da rede relativos àquela escolha de núcleo. No caso de uma rede de função de base radial, o núcleo é uma função gaussiana. Para este método de implementação, o número de funções de base radial e seus centros e seus pesos são calculados automaticamente. Os centros das funções de base radial são definidos pelos vetores de suporte escolhidos pela estratégia de otimização quadrática. Os vetores de suporte são tipicamente uma fração do número total de exemplos que constituem a amostra de treinamento [29].

2.4 Utilização do SVM para calibração multivariada

Como em um modelo SVM para classificação, os dados originais são mapeados em um espaço de alta dimensão e, em seguida, uma função linear é adequada para aproximar a função latente entre **X** (matriz de dados) e **y** (vetor resposta).

A fim de usar o SVM para calibração foi proposto [35] que se transformasse o problema de calibração em um problema de classificação. Para cada amostra x_i do conjunto de treinamento um y_i correspondente é adicionado a um número positivo d para produzir uma nova amostra (x_i,y_i^{1}) pertencente a classe 1. De forma similar, o y_i pode também ser subtraído pelo mesmo d para produzir outra nova amostra (x_i,y_i^{-1}) pertencente a classe -1. Repetindo esse processo, as N amostras para calibração são duplicadas e dispostas em duas classes, assim a calibração é transformada em um problema de classificação binário, como mostrado na Figura 9A. Outro parâmetro importante para os modelos SVM usados para calibração é *a ϵ-band,* região definida como na Figura 9B [30].

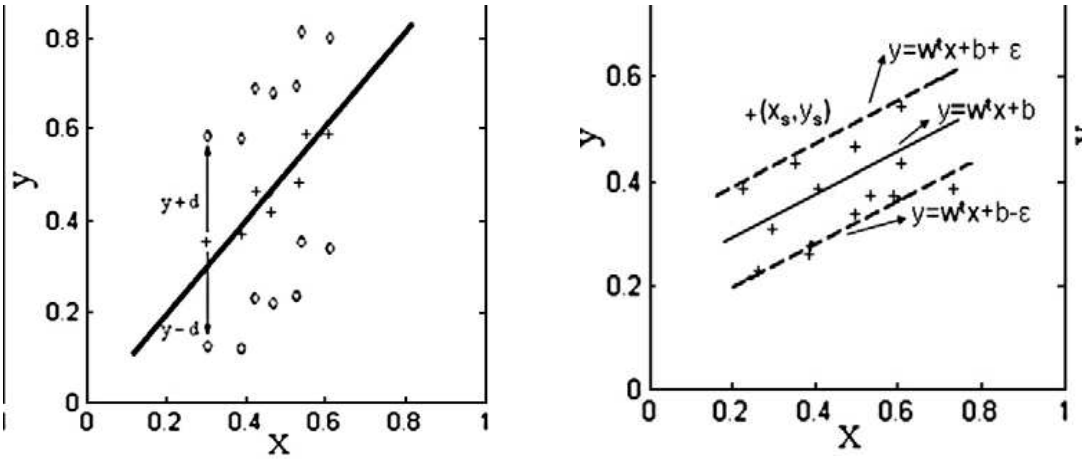


Figura 9. A) Transformação de um problema de calibração em um de classificação. **B)** ϵ-*band*

Tendo em vista a construção de um modelo de calibração robusto, deve haver a preocupação com a máxima degradação de desempenho que é possível para um desvio ϵ (Figura 9B). Para resolver o problema de regressão não-linear o

erro absoluto é usado como quantidade a ser minimizada, assim a função de perda tem a forma:

$$L(d,y) = |d - y| \qquad \text{Eq. 53}$$

onde d é a resposta desejada e y a saída do estimador.

Para construir um modelo SVM para aproximar uma resposta desejada d, pode-se utilizar a extensão da função de perda, como descrito [29]:

$$L_\epsilon(d,y) = |d - y| - \epsilon, \text{ para } |d - y| \geq \epsilon \qquad \text{Eq. 54}$$

$$L_\epsilon(d,y) = 0, \text{ caso contrário} \qquad \text{Eq. 55}$$

onde ϵ é um parâmetro predeterminado. A função de perda é chamada de função de perda insensível a ϵ. Ela é igual a zero se o valor absoluto do desvio da saída do estimador y em relação a resposta desejada d for menor ou igual a zero, caso contrário, ela é igual ao valor absoluto do desvio menos ϵ.

A Figura 10 [30] ilustra a dependência de $L_\epsilon(d,y)$ em relação ao erro d – y [29]. Isso quer dizer, apenas os dados fora da ϵ-*band* (região entre as linhas pontilhadas da Figura 9B) causam perda. A função de perda da Eq.53 é um caso especial da função de perda insensível a ϵ para ϵ=0.

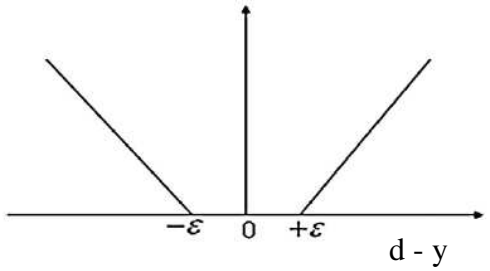


Figura 10. Curva da função de perda insensível a ϵ.

A partir de um conjunto de dados $D = \{(x_i, y_i)\}_{i=1}^{N}$ o SVM para calibração de dados lineares será escrito de forma a seguir com as funções de perda, nos seguintes termos [30]:

$$f(\mathbf{w},\varepsilon) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{N}\sum_{i=1}^{N} L\left(y_i - f(x_i),\varepsilon\right) \qquad \text{Eq. 56}$$

onde γ é o parâmetro de regularização pré-definido e ξi(*) é introduzido como variável "solta" para definir a superfície de decisão.

$$L\left(w,b,\xi^{(*)}\right) = \frac{1}{2}\|w\|^2 + \frac{\gamma}{N_i}\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \qquad \text{Eq. 57}$$

Assim como no SVM utilizado para classificação essa otimização restrita, é resolvida através dos multiplicadores de Lagrange. A função de decisão será dada por:

$$f(\mathbf{x},\mathbf{w}) = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)K(x_i,x) + b, \qquad \text{Eq. 58}$$

sendo que os termos α_i e α_i^* otimizados a partir de multiplicadores de Lagrange, K a função de kernel aplicada e b representando os desvios "bias" [30].

2.5 Utilização do algoritmo genético para otimização dos parâmetros do SVM

O GA evolucionário implementado segundo Huang [36] otimiza os dois parâmetros utilizados no SVM, γ e δ^2 seguindo uma arquitetura como a representada na Figura 11.

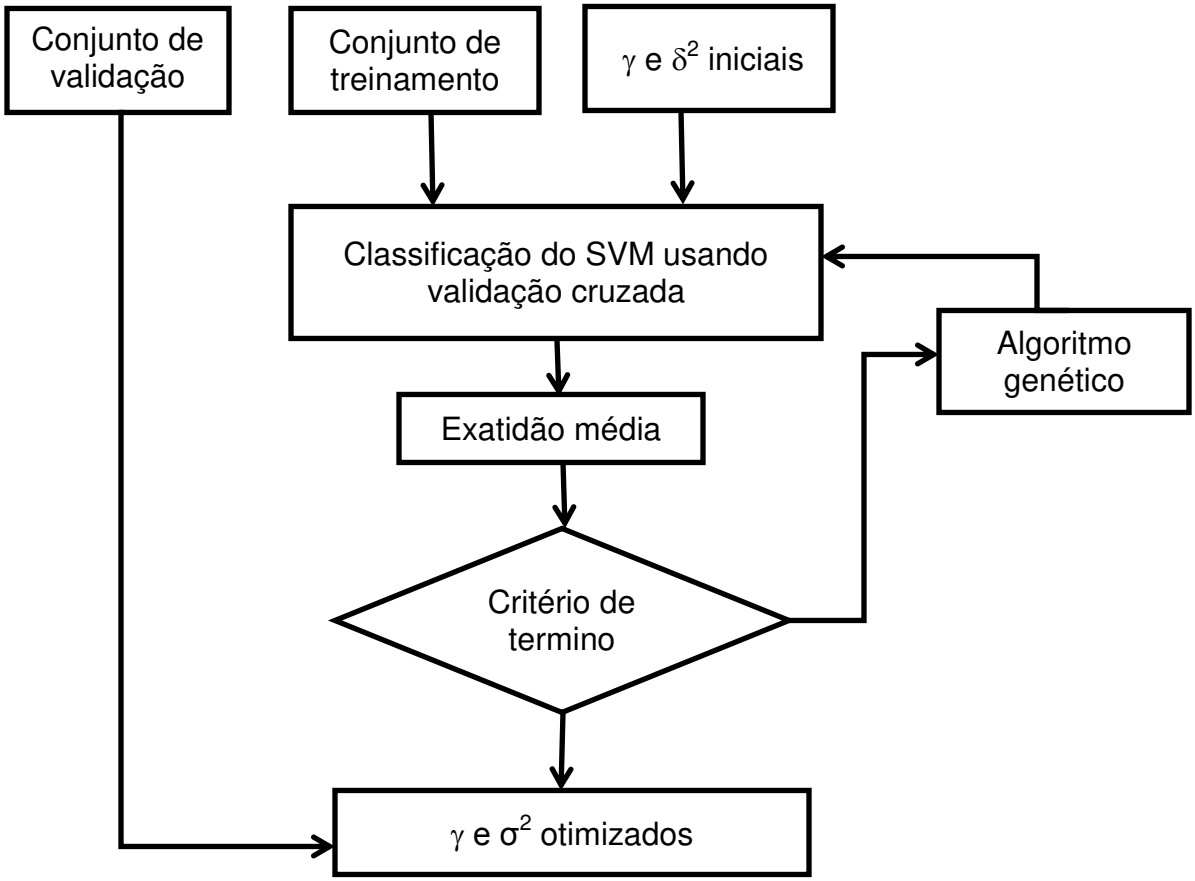


Figura 11. GA aplicado a SVM

A arquitetura proposta para o algoritmo genético é dada a seguir:

(1) Escalonamento dos dados (pré-processamento), com o propósito de aumentar a exatidão do SVM. Geralmente, os dados podem ser linearmente escalonados em [-1,+1] ou [0,1] através da seguinte equação:

$$\nu' = \frac{\nu - \min_a}{\max_a - \min_a} \qquad \text{Eq. 59}$$

(2) Conversão do genótipo em fenótipo. Durante esta etapa, cada parâmetro e característica do cromossomo são transferidos.

(3) Após a aplicação do algoritmo genético que converteu cada característica genética do subconjunto para o novo fenótipo, a característica do subconjunto poderá ser determinada.

(4) A avaliação da resposta de cada cromossomo representado por γ e σ^2, será avaliada para a matriz de treinamento e para a matriz de validação, em relação ao erro médio quadrático (MSE - *mean square error*) e ao coeficiente de determinação entre valores esperados e previstos.

(5) Critério de término. Quando o critério de término é satisfeito, o processo termina; se não, a próxima geração é processada.

(6) Operação genética. Neste passo, o sistema procura pela melhor solução por operações genéticas, incluindo seleção, cruzamento, mutação e substituição.

Nesta tese o critério de término das gerações foi a estabilização dos resultados da validação cruzada, de forma que eram produzidas novas gerações até que não houvessem mais diferenças significativas dos resultados quando a nova geração é comparada à anterior.

Capítulo III

3. Determinação de Carbono e Nitrogênio em solo por NIR

As mudanças no uso e cobertura do solo, principalmente na região tropical, têm grande importância para o ciclo global do Carbono e do Nitrogênio, pois a ciclagem da matéria orgânica do solo é mais intensa em ecossistemas de clima tropical do que em clima temperado e tende a diminuir com o uso agrícola. Um exemplo marcante de mudança no uso e cobertura do solo é a substituição de florestas por pastagens e cultivos agrícolas no sudoeste da Amazônia, principalmente nos Estados de Rondônia e Mato Grosso [37].

No entanto, para que se possam sugerir alternativas de manejo viáveis para ecossistemas complexos como o dos solos tropicais do Brasil, faz-se necessário a realização de estudos que avaliem de forma integrada as informações sobre o solo, o clima e a vegetação da região como um todo. Para alcançar este objetivo faz-se necessário o uso de técnicas analíticas rápidas, precisas e que possam ser aplicadas no campo, associadas a análises estatísticas, uma vez que a quantidade de amostras e dados gerados é muito alta. Atualmente há um grande número de trabalhos científicos reportando o uso de espectroscopia NIR para analises de propriedades de solos [38-42].

Assim, o primeiro caso estudado nesta tese foi a construção de modelos de calibração multivariada e espectroscopia no infravermelho próximo a fim de prever quantidades de carbono e nitrogênio em solo contendo ou não o mineral gipsita.

A gipsita é um mineral abundante na natureza, é um sulfato de cálcio hidratado cuja fórmula química é $CaSO_4.2H_2O$, que geralmente ocorre associado à anidrita e tem pouca expressão econômica [43].

Na agricultura, moída na granulometria apropriada, a gipsita é utilizada como corretivo do pH de solos, tendo sua aplicação se dado inicialmente na Europa, nos primórdios do século XVIII. A partir daí vem sendo cada vez mais utilizada na correção de solos alcalinos onde, ao reagir com o carbonato de sódio, dá origem ao carbonato de cálcio e o sulfato de sódio, substâncias de grande importância agrícola. Também é utilizada como corretivo de solos deficientes em

enxofre, para possibilitar a assimilação do potássio e o aumento do conteúdo de nitrogênio [43].

3.1 Experimental

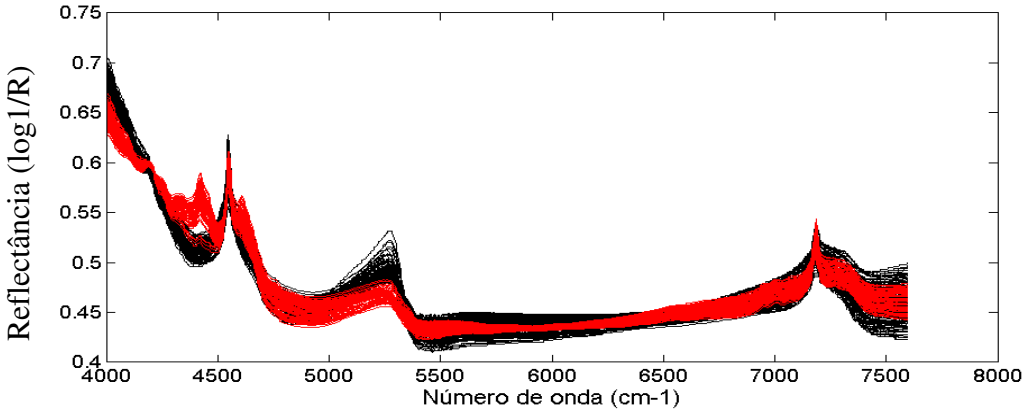
As amostras foram colhidas em um campo experimental de cana-de-açúcar cultivada sob colheita mecanizada e sem queima. Elas estavam localizadas na cidade de Pradópolis, no estado de São Paulo em talhões nos quais a cana havia sido plantada há 2, 4, 6 e 8 anos, sem ter havido reforma do canavial, e em uma área de vegetação nativa. A amostragem foi realizada em nove pontos, nas profundidades 0-10, 10-20, 20-30, 40-50, 70-80 e 90-100 cm, totalizando 203 amostras.

As amostras foram secas, peneiradas e moídas para ficarem na granulometria de 60 mesh. O método de referência para Carbono e Nitrogênio totais foi a combustão a seco em um analisador elementar LECO CN 2000. O princípio do método é a conversão de todas as diferentes formas de Carbono a CO_2, que pode ser medido quantitativamente por infravermelho. Adicionalmente todo Nitrogênio passa por catalisadores e filtros, para ser então detectado (na forma N_2) por um detector termoelétrico. Um padrão de solo da marca LECO foi utilizado para a construção da curva de calibração interna para C e N e aferição diária do auto-analisador, em termos de repetibilidade e precisão. Cada amostra foi analisada em triplicata, com erro relativo inferior a 5%. A concentração de Carbono variava entre 0,35 e 4,80% e a concentração de Nitrogênio entre 0,038 e 0,32 % em massa.

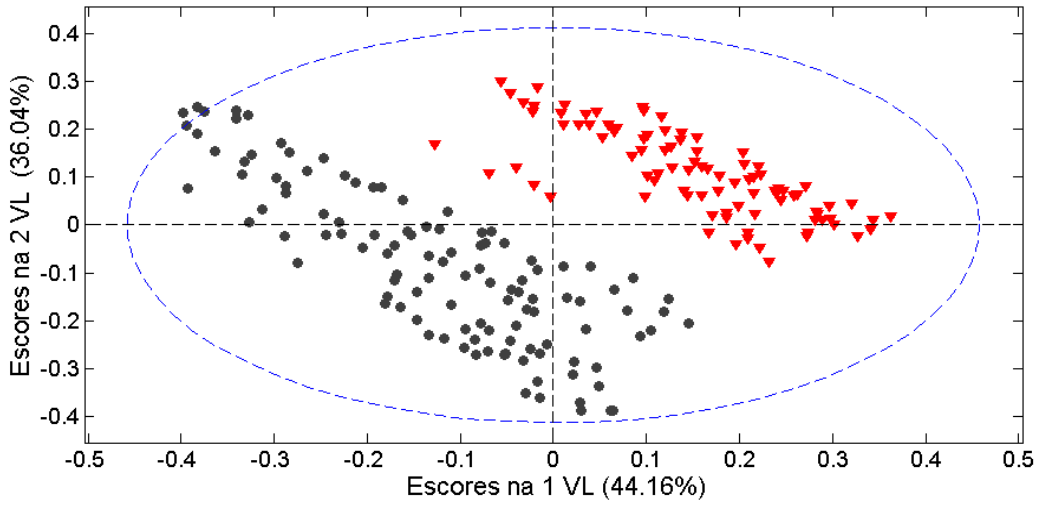
Os espectros das amostras de solo foram obtidos na região do infravermelho próximo (4000 a 7600 cm^{-1}) em intervalos de 2 cm^{-1} em um equipamento NIRS 5000 da Foss NIRSystems. Todos esses procedimentos experimentais foram realizados na ESALQ de Piracicaba. Para tratamento dos dados e construção dos modelos quimiométricos foram utilizados o Matlab 7.0.1, PLS-Toolbox 4.02 e a rotina de SVM proposta por Pelckmans *et all* [44].

3.2 Resultados e discussões

Os espectros foram tratados com correção multiplicativa de sinal (MSC, *"Multiplicative Scatter Correction"*) para eliminar problemas de espalhamento de radiação. Na Figura 12A estão todos os espectros coletados. Os espectros em vermelho são referentes as 91 amostras de solo com gipsita e os em preto se referem as 112 amostras de solo sem gipsita.



A



B

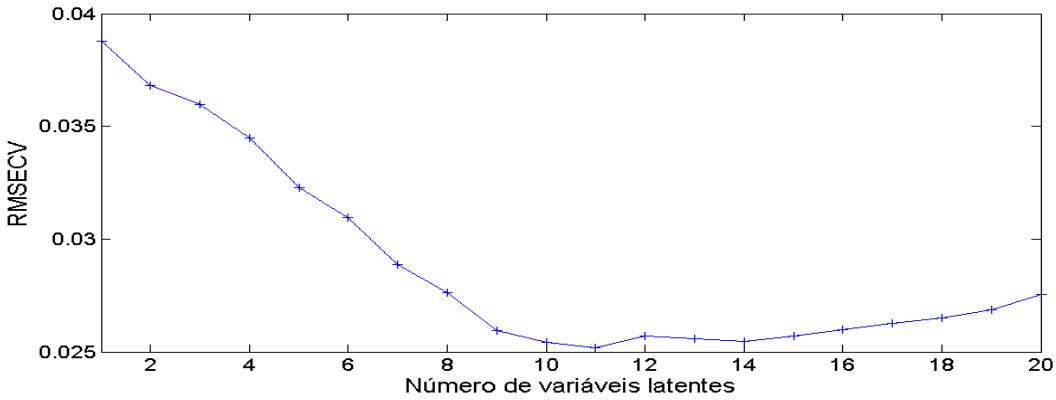
Figura 12. A) Espectros de infravermelho próximo das amostras de solo e **B)** Escores das 1ª e 2ª variáveis latentes. As amostras de solo com gipsita estão representadas em vermelho e as amostras de solo sem gipsita estão representadas em preto.

Na Figura 12A observa-se a diferença entre os espectros das amostras de solo com e sem gipsita. Para evidenciar essa diferença, foi construído um modelo de PCA usando os espectros centrados na média. Os escores das duas primeiras variáveis latentes estão na Figura 12B, onde amostras de solo com gipsita estão representadas em vermelho e as amostras de solo sem gipsita estão representadas em preto. É nítido que os dois tipos de solo formam agrupamentos distintos, conferindo um problema adicional na quantificação de Carbono e Nitrogênio por NIR. A proposta do trabalho foi construir um modelo de previsão para esses elementos que seja independente da presença ou não de Gipsita no solo, já que em alguns casos pode não haver essa informação para o analista.

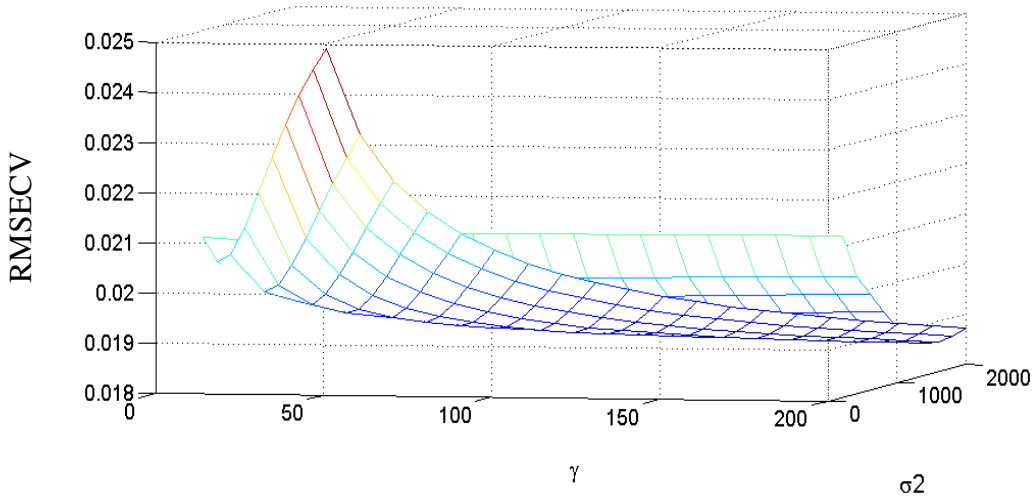
Foram construídos modelos PLS e SVM para os solos com e sem gipsita separadamente e também usando todas as amostras, independentemente do tipo de solo. Para construção de todos os modelos as amostras foram separadas em conjuntos de calibração e validação através do algoritmo de Kennard-Stone [45] com 45 amostras de solo com guipsita e 57 amostras de solo sem guipsita ficando no conjunto de validação.

Os modelos de PLS foram construídos utilizando os conjuntos de calibração, com os espectros centrados na média. As variáveis latentes de cada modelo foram escolhidas através da validação cruzada "*leave one out*" como no exemplo da Figura 13A, onde foram escolhidas aquelas que geravam modelos com menor RMSECV.

Foram preparados também os modelos SVM utilizando os mesmos conjuntos de calibração. Os parâmetros γ e σ^2 foram escolhidos empiricamente através da observação da superfície de resposta destes parâmetros contra valores de RMSECV, como na Figura 13B. A partir desta figura foram escolhidos os valores dos parâmetros com menor RMSECV associado, ou seja, que geram o modelo mais ajustado.



A



B

Figura 13. A) Exemplo de gráfico de RMSECV por número de variáveis latentes e **B)** Superfície de RMSECV por γ e σ^2.

Com os parâmetros ótimos escolhidos através da observação da superfície de RMSECV por γ e σ^2 e modelos PLS e SVM construídos foram previstos os valores de Carbono e Nitrogênio nas amostras de validação através de ambos os algoritmos.

A Figura 14 apresenta os gráficos de valores reais contra valores previstos dos modelos SVM e PLS quando estes foram construídos utilizando todos os espectros, enquanto na Figura 15 observa-se estes mesmos gráficos quando gerados por modelos que utilizaram apenas espectros de amostras com ou sem gipsita.

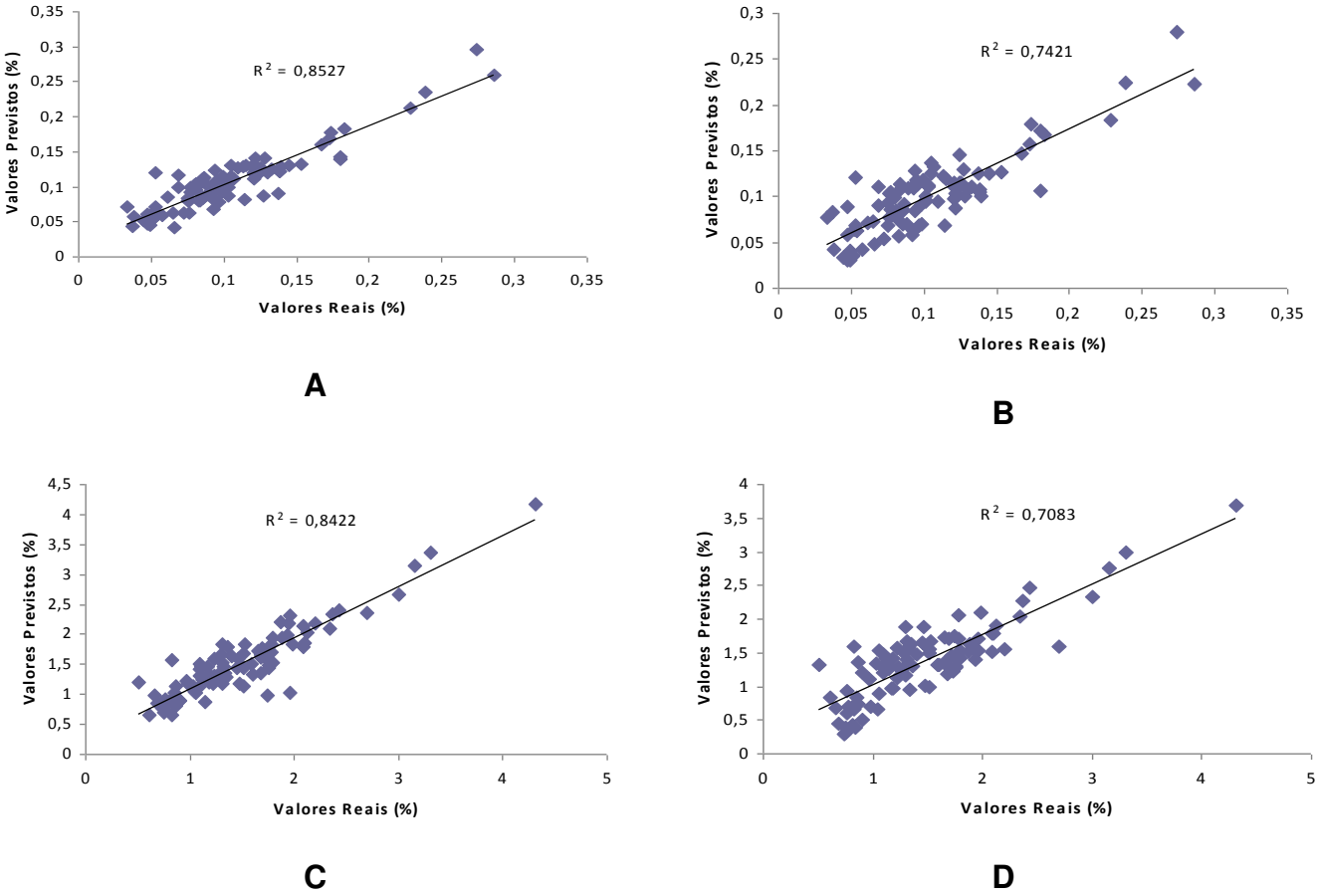


Figura 14. Valores reais contra valores previstos de Nitrogênio em solo para modelos **A)** SVM e **B)** PLS, e de Carbono para modelos **C)** SVM e **D)** PLS

Nos gráficos da Figura 14 podemos ver que tanto para Carbono quanto para Nitrogênio, os modelos SVM apresentaram valores mais próximos dos reais do que os modelos PLS, o que pode ser observado através da distribuição dos pontos ao longo da reta e do coeficiente de determinação de cada reta.

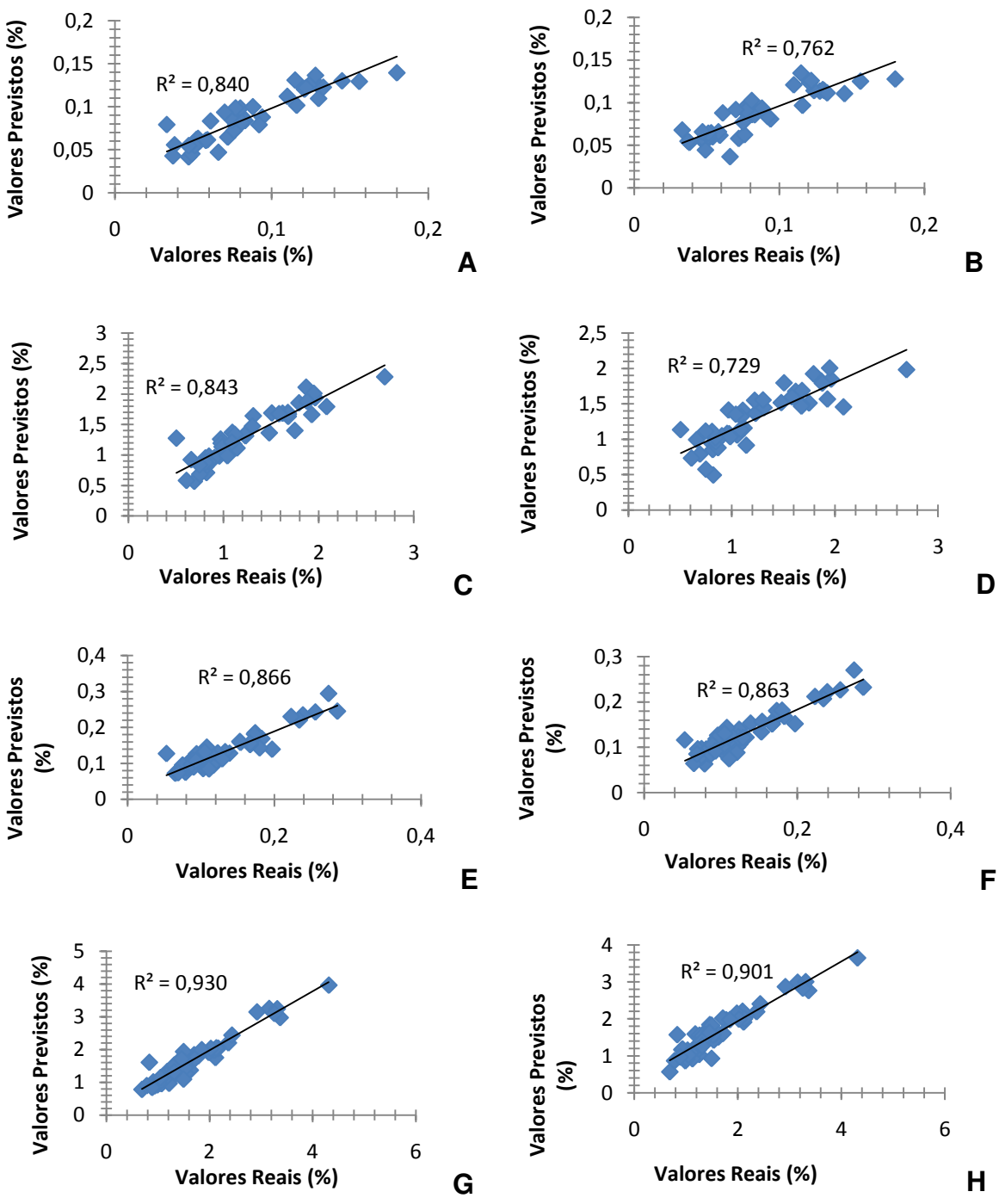


Figura 15. Valores reais contra valores previstos de Nitrogênio em solo com gipsita para modelos **A)** SVM e **B)** PLS, de Carbono em solo com gipsita para modelos **C)** SVM e **D)** PLS; de Nitrogênio em solo sem gipsita para modelos **E)** SVM e **F)** PLS; de Carbono em solo sem gipsita para modelos **G)** SVM e **H)** PLS.

Na Figura 15 pode-se notar que o SVM produziu previsões mais próximas dos valores reais nos modelos construídos com amostras de solos com gipsita, tanto para Carbono quanto para Nitrogênio, entretanto, nas previsões para solos sem gipsita as relações foram praticamente idênticas. Os coeficientes de determinação para todos os gráficos de valores reais contra valores previstos estão na Tabela 2.

Tabela 2. Coeficiente de determinação para os gráficos de valores reais contra valores previstos dos modelos SVM e PLS construídos.

		Nitrogênio	**Carbono**
*Mistura de solos**	**PLS**	0,742	0,708
	SVM	0,852	0,842
Solo com gipsita	**PLS**	0,762	0,729
	SVM	0,840	0,843
Solo sem gipsita	**PLS**	0,863	0,901
	SVM	0,866	0,930

* Amostras de solo com e sem gipsita

Outra forma de avaliar a diferença entre os modelos é através da Raiz Quadrada do Erro Médio Quadrático de Previsão. Os RMSEPs mostram o erro global das previsões e, para confirmação estatística da diferença de desempenho entre os modelos, pode ser realizado um teste F. Esse teste F é feito utilizando os valores dos RMSEPs obtidos pelos algoritmos empregados para cada situação proposta, onde o valor do F tabelado para a mistura de solos foi para 95 graus de liberdade para os modelos feitos com a mistura de solos, 45 graus de liberdade para os modelos feitos com solos com gipsita e 56 graus de liberdade para os modelos feitos com solos sem gipsita. Os RMSEPs de todos os modelos obtidos, assim como os valores de F calculado e valores de F tabelados estão na Tabela 3.

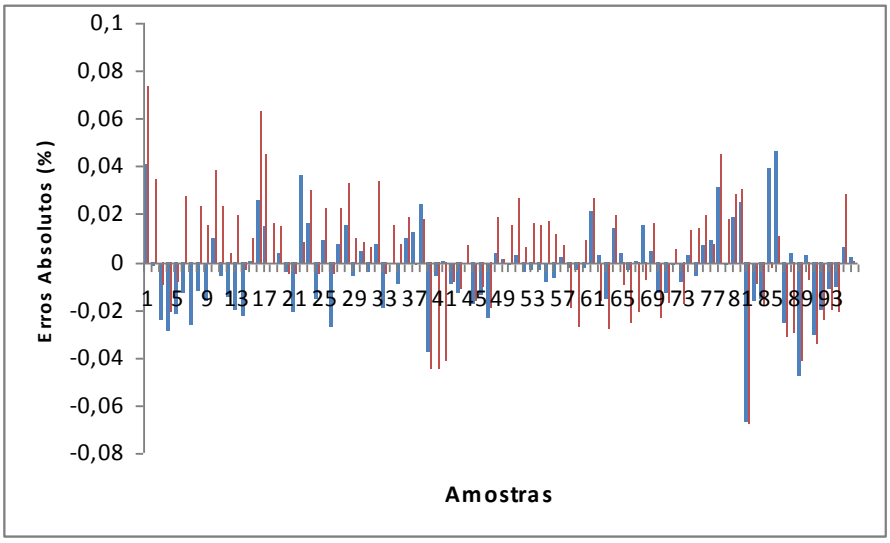
Tabela 3. Resultados dos modelos obtidos

Nitrogênio		Mistura de solos* (%)	Solo com gipsita (%)	Solo sem gipsita (%)
RMSEP	**PLS**	0,024	0,018	0,021
	SVM	0,018	0,015	0,020
	F calculado	1,74	1,49	1,07
	F Tabelado (95%)	1,15	1,59	1,35
Carbono		**Mistura de solos* (%)**	**Solo com gipsita (%)**	**Solo sem gipsita (%)**
RMSEP	**PLS**	0,12	0,26	0,24
	SVM	0,064	0,20	0,20
	F calculado	1,96	1,29	1,23
	F Tabelado (95%)	1,15	1,59	1,35

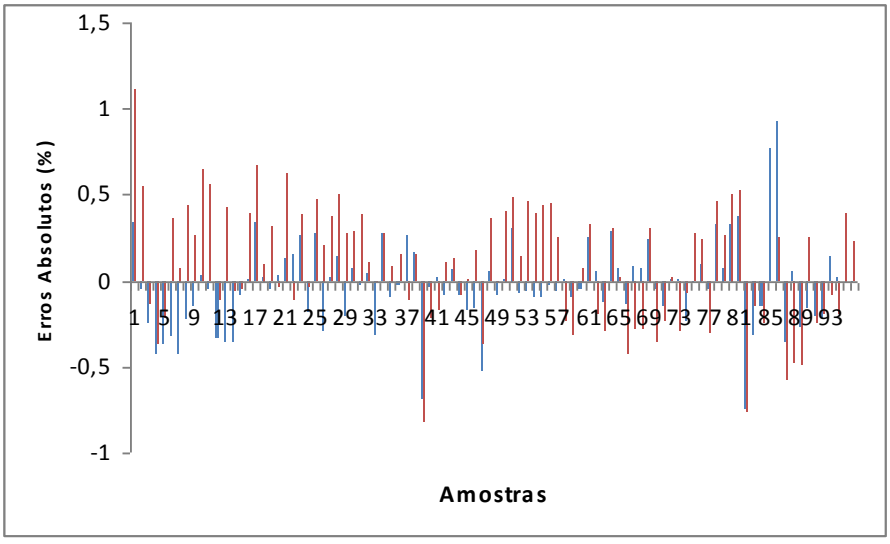
* Amostras de solo com e sem gipsita

Na Tabela 3 pode-se ver que o SVM produziu RMSEPs menores do que o PLS em todos os seis casos, entretanto, através do teste F pode-se afirmar com 95% de certeza que apenas nos modelos construídos com todas as amostras de solo, ou seja, na presença ou ausência do mineral gipsita, o SVM produziu amostras com erros de previsão consideravelmente menores do que o PLS. Nos modelos construídos com as amostras de solo separadas, não houve indício estatístico de que um algoritmo produziu modelos melhores do que o outro.

Por fim, foram construídos gráficos mostrando os erros absolutos obtidos com os modelos PLS e SVM das amostras de solo com a mistura (Figura 16).



A



B

Figura 16. Erros absolutos para **A)** Nitrogênio e **B)** Carbono. Em azul estão os erros de previsão dos modelos SVM e em vermelho os erros de previsão dos modelos PLS.

A Figura 16 mostra que os erros absolutos do PLS foram maiores que os do SVM na maior parte das previsões das concentrações de Nitrogênio e Carbono em solo, independentemente da faixa de concentração dos mesmos.

3.3 Conclusões

Com base no teste F, nos gráficos de valores reais contra previstos e nos erros absolutos, pode-se concluir que o desempenho do SVM na previsão da concentração de Nitrogênio e Carbono nas amostras de solo, sem a discriminação dos mesmos quanto à presença do mineral guipsita é melhor do que do PLS, produzindo erros de previsão significantemente menores.

Quando os tipos de solos são separados o SVM deixa de ter um desempenho melhor do que o PLS, sendo equivalentes, já que o conjunto de amostras perde a característica de separação em grupos distintos de amostras. Ou seja, o SVM se mostrou um algoritmo mais apropriado, em comparação ao PLS, para problemas de calibração multivariada onde há diferentes classes de amostras.

Capítulo IV

4. Determinação de Mineral em Polímero por NIR

O maior desafio na análise de misturas sólidas está no processo de amostragem, o qual deve ser representativo e não exercer nenhum tipo de influência na composição do produto final a ser analisado. O método tradicional de amostragem consiste em remover porções de diferentes pontos da mistura, entretanto, essa metodologia pode gerar vários problemas como segregação da mistura, quantidade limitada de material que pode ser utilizada e alteração da composição da mistura, entre outros [46-49], portanto, devido a uma série de vantagens, a espectroscopia na região do infravermelho próximo tem despertado o interesse de diversos setores industriais como ferramenta para o controle de qualidade. Esse interesse decorre do fato dessa técnica permitir o controle de qualidade de misturas sólidas, possibilitando a caracterização, otimização e o controle de processos de materiais sólidos em tempo real [50-53].

O objetivo desta aplicação foi desenvolver modelos de calibração multivariada a partir de dados de espectroscopia no infravermelho próximo que conseguisse prever a concentração de um mineral adicionado a um polímero sem a necessidade de abertura da amostra e a comparação entre os algoritmos usados para a construção desses modelos, sendo esses o SVM e o PLS. As propriedades físicas e químicas deste polímero estão intimamente ligadas com a concentração do mineral durante seu processo de fabricação. Portanto, o controle da concentração desse mineral é extremamente importante para a qualidade do produto.

Um conjunto de 54 espectros de reflectância difusa de um determinado polímero, com concentração de um mineral entre 0,00 e 1,63% em massa foi fornecido pela empresa Rhodia, sediada em Lion (França). Por ser um trabalho em cooperação com uma empresa, não tivemos acesso à composição ou qualquer outra informação física ou química do polímero ou do mineral. Acreditamos que para o objetivo deste trabalho essas informações não são de importância fundamental, pois se deseja apenas avaliar os diferentes modelos de calibração multivariada.

Os espectros foram obtidos em um espectrofotômetro da marca NIR-System Foss 6500 acoplado a um acessório de reflectância difusa na faixa entre 1100 e 2500 nm, com resolução de 2 nm e 32 leituras por espectro. As amostras tiveram as porcentagens de mineral estimadas através de um método padrão de análise, que foi através da técnica de absorção atômica, com desvios de aproximadamente ± 0,008%. Para tratamento dos dados e construção dos modelos quimiométricos foram utilizados o Matlab 7.0.1, PLS-Toolbox 4.02 e a rotina de SVM proposta por Pelckmans *et all* [44].

Para minimizar o desvio da linha de base ocasionado pelo espalhamento de luz das amostras foi utilizado o pré-processamento Transformação Padrão Normal de Variação (SNV). Essa ferramenta apresenta bom desempenho na eliminação do efeito do espalhamento de luz multiplicativo e, além disso, apresenta a vantagem de não necessitar da informação de outros espectros para realizar a correção da linha base, podendo ser empregada em sistemas onde o perfil dos espectros possui alta variabilidade. Os espectros foram também centrados na média. Os espectros antes (Figura 17) e depois dos pré-tratamentos (Figura 18) são mostrados a seguir:

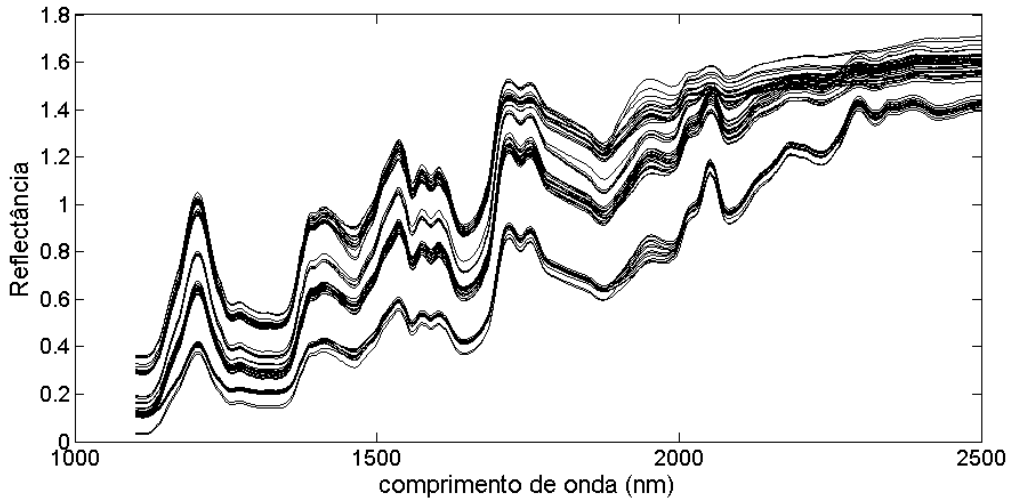


Figura 17. Espectros das amostras de polímeros sem pré-tratamento.

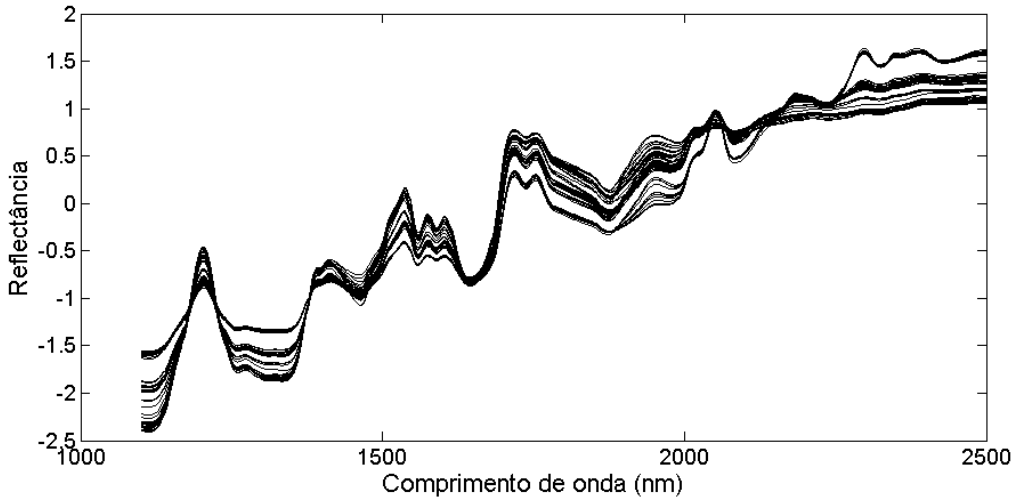


Figura 18. Espectros das amostras de polímeros após pré-tratamentos.

As amostras foram divididas em grupos de calibração (com 30 amostras) e validação (com 24 amostras) através do algoritmo de Kennard-Stone. Foi feito então uma PCA com as amostras de calibração. Os escores podem ser vistos na Figura 19 e as concentrações das amostras estão na Tabela 4:

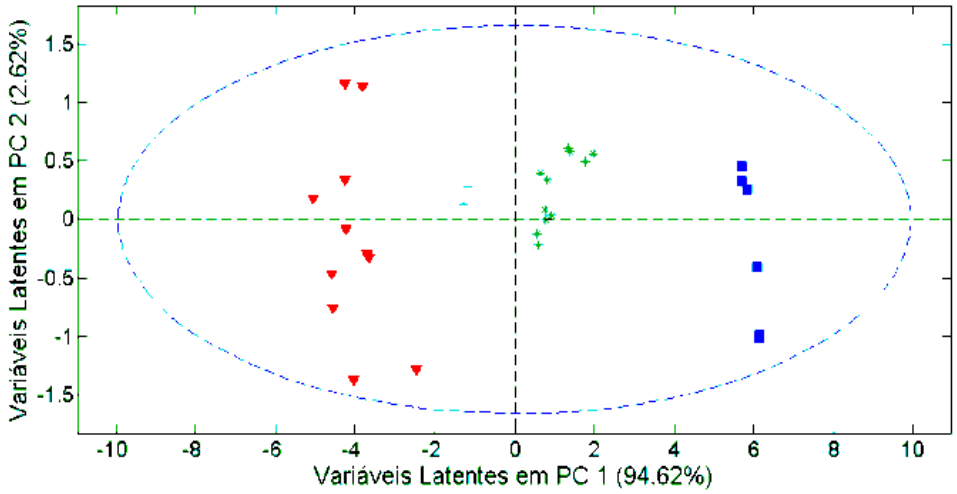


Figura 19. Primeira e Segunda componentes principais do modelo PCA.

Na Figura 20 os espectros estão apresentados com as mesmas cores dos grupos do PCA.

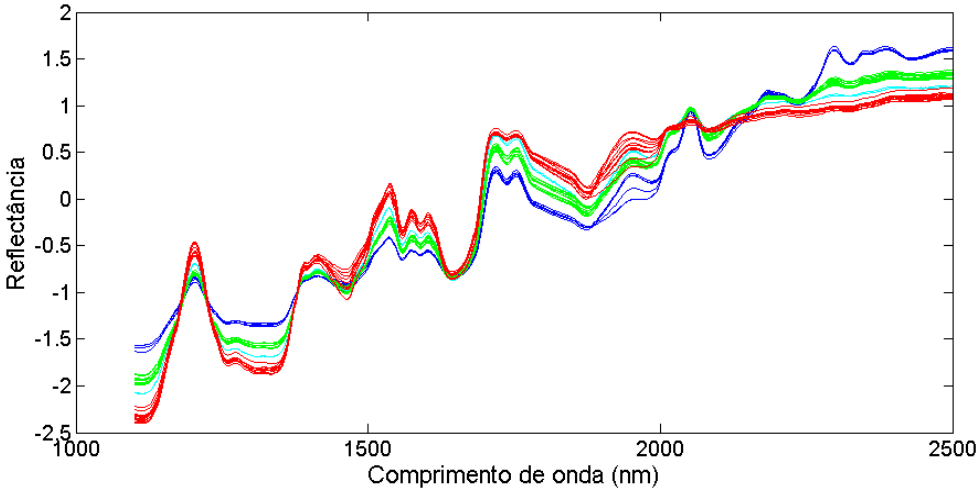


Figura 20. Espectros separados em cores por diferentes faixas de concentração.

Tabela 4. Concentração das amostras de calibração

Conjunto	Concentração do mineral (%)
▼	0 – 0,021
+	0,105 – 0,125
*	0,291 – 0,320
■	1,631 – 1,750

É possível observar através das Figuras 19 e 20 e da Tabela 4, que há uma diferenciação das amostras em grupos, dependendo da faixa de concentração do mineral.

Através da Figura 20, podemos notar variações significativas no perfil dos espectros em função da concentração do mineral no polímero, o que sugere correlação entre intensidade do espectro e concentração do mineral. A maior variação se dá em comprimentos de onda maiores do que 2000 nm, região onde espectros relacionados a amostras sem o mineral têm pouca definição de bandas e pequenas alterações da concentração do mesmo causam grandes aumentos de picos. Segundo a literatura [40], bandas ao redor de 1100, 1500, 1900, 2050 e 2250 nm podem ser associadas às vibrações das ligações químicas N-H, C-H, O-H e C-O. Portanto, uma ou mais ligações desses grupos orgânicos poderiam estar

interagindo diretamente com o mineral, alterando propriedades físicas e químicas do polímero.

Essa variação dos espectros é confirmada através do PCA construído com essas amostras, como podemos ver na Figura 19, onde há uma clara separação de grupos em PC1, que são determinados pelo aumento da concentração do mineral.

4.1 Modelo PLS

Foi construído um modelo PLS para previsão dos dados. Na construção do modelo foi utilizado o conjunto de calibração e para a escolha do número de variáveis latentes foi utilizada a validação cruzada *"leave one out*", que obteve como resultado a Figura 21.

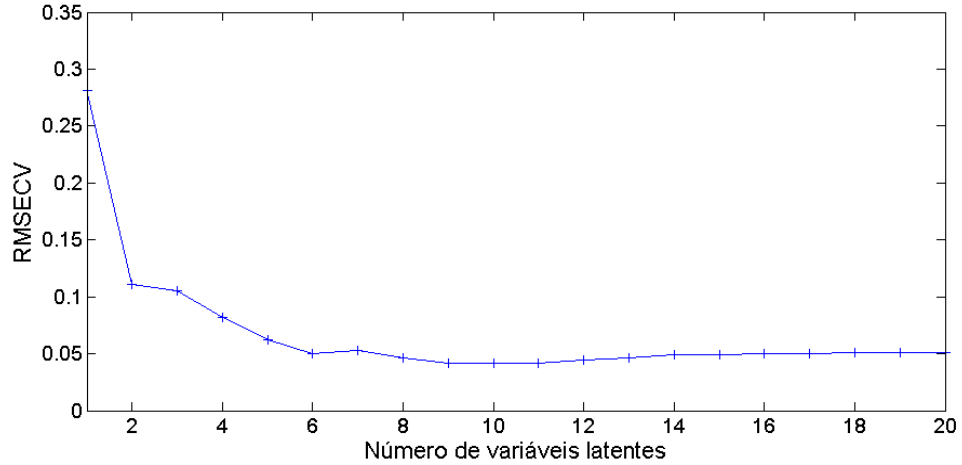


Figura 21. RMSECV por número de variáveis latentes.

O modelo foi construído com seis variáveis latentes e então foi feita a previsão das amostras de validação, obtendo um RMSEP de 0,056%. Esse modelo obteve coeficiente de determinação entre os valores reais e previstos de 0,995, como mostra a Figura 22.

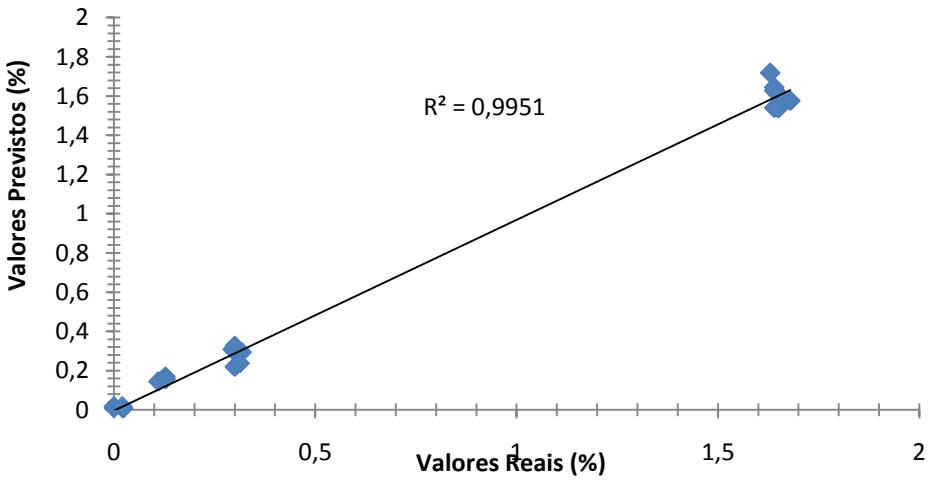


Figura 22. Valores reais contra valores previstos no modelo PLS.

Os resultados obtidos mostram que apesar de os conjuntos de dados terem distinções suficientes para formarem diferentes grupos no PCA, o PLS foi capaz de produzir resultados satisfatórios.

4.2 Modelo SVM

Foi desenvolvido então um modelo utilizando o SVM. Para otimizar os parâmetros do algoritmo foi utilizada a superfície de RMSECV por γ e σ^2, sendo escolhidos como parâmetros ótimos γ =200 e σ^2 =300, como mostra a Figura 23.

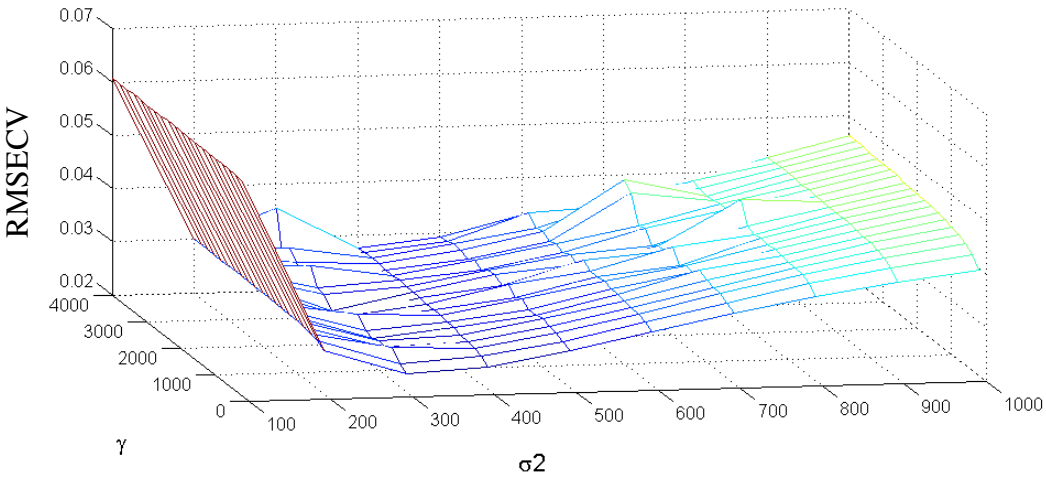


Figura 23. Superfície de RMSECV por γ e σ^2

Foram feitas as previsões das amostras de validação, usando este modelo. O RMSEP deste modelo foi de 0,023% e o coeficiente de determinação entre valores reais e previstos foi de 0,999, como mostra a Figura 24.

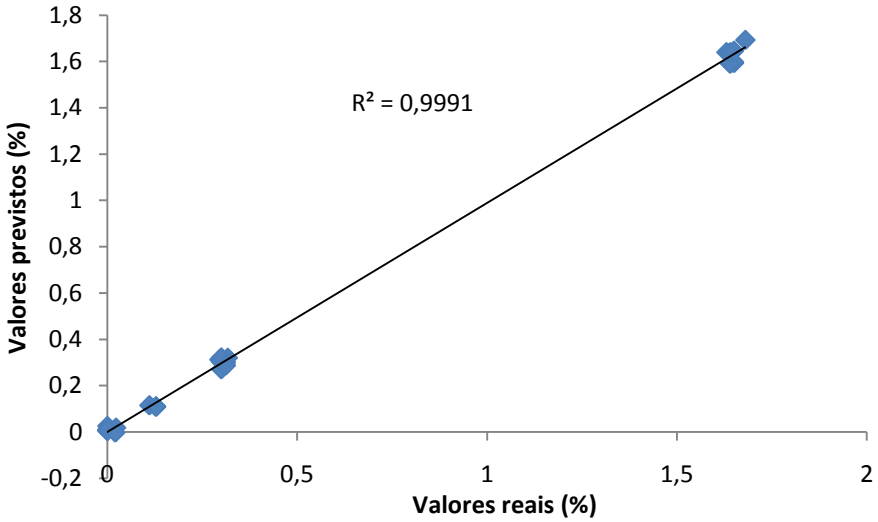


Figura 24. Valores reais contra valores previstos no modelo SVM.

Os erros absolutos das previsões para ambos os algoritmos, PLS e SVM, estão apresentados na Figura 25.

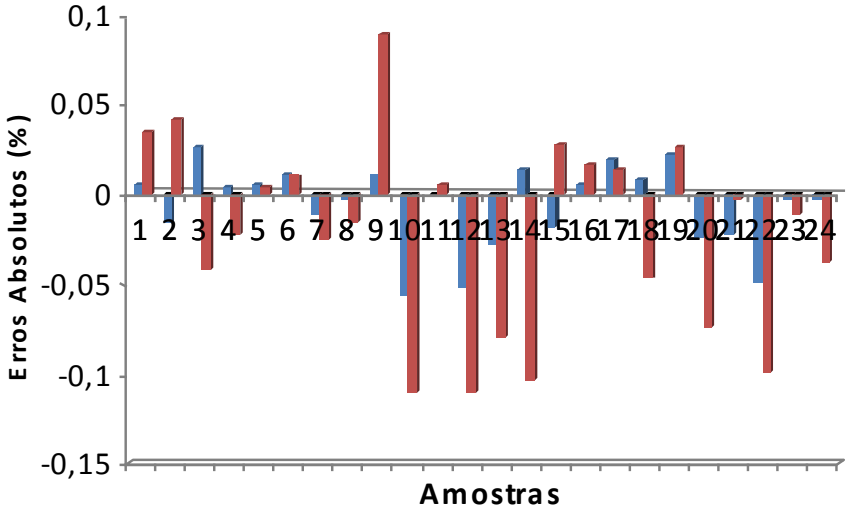


Figura 25. Erros absolutos de previsão de modelos PLS (em vermelho) e SVM (em azul) para porcentagem de minério em polímero.

4.3 Conclusões

Tanto o coeficiente de determinação calculado entre valores esperados e os calculados pelos modelos quanto o RMSEP do modelo do SVM foram melhores do que do PLS. Para comparação estatística dos métodos foi feito um teste F com os RMSEPS obtidos em ambos.

O F calculado foi 5,92 enquanto o F tabelado, para 23 graus de liberdade e grau de confiança de 95% é aproximadamente 2,01. Assim, é possível afirmar estatisticamente que o SVM teve um desempenho diferenciado em relação ao PLS na previsão das porcentagens de minério nas amostras de polímero, neste caso, produzindo previsões com erros consideravelmente menores.

Capítulo V

5. Diferenciação de bactérias quanto à condição de crescimento por MIR

Acidithiobacillus ferrooxidans é uma bactéria Gram-negativa acidófila cujas condições ótimas de crescimento são pH em torno de 2,0 e temperatura de 30 º C. Ela obtém energia a partir da oxidação do ferro ou compostos reduzidos de enxofre e é usada industrialmente na lixiviação de metais, processo no qual sulfetos de metal são convertidos em sulfatos de metal solúveis em água [54]. Durante o processo de biolixiviação, a *A. ferrooxidans* é frequentemente sujeita a mudanças no pH, temperatura ideal de crescimento e de nutrientes [54]. Essas mudanças podem afetar a fisiologia de bactérias e, como consequência, a eficiência da biolixiviação.

A *Acidithiobacillus ferrooxidans* responde às altas temperaturas pela síntese de diversas proteínas de choque térmico [55,56]. Esta bactéria também é capaz de adquirir termotolerância [57], que indica que ela desenvolveu mecanismos de proteção para lidar com o estresse de calor. Este fato é particularmente importante já que a temperatura é um dos principais fatores que afetam a solubilização de metais durante a lixiviação [58].

Quanto à privação de fosfato, foi demonstrada redução na taxa de crescimento da bactéria, bem como na sua capacidade de oxidar o ferro ferroso e fixar CO_2 [59]. Também já foi observado aumento da fosforilação de proteínas das células, sugerindo a ativação de uma resposta ao estresse geral [60].

Além disso, a produção de lipopolissacarídeos nas células de *A. ferrooxidans* é alterada na privação de fosfato [61], o que pode afetar a biolixiviação, já que lipopolissacarídeos são parte da matriz de polissacarídeo envolvidos na colonização do minério.

A importância do processo de biolixiviação para a economia mundial e para preservação do meio ambiente ocorre em virtude do aumento da demanda mundial de bens minerais, o que tem provocado esgotamento progressivo de reservas contendo altos teores de metais de interesse econômico. Assim, a aplicação da biolixiviação para recuperação de metais a partir de minérios de baixos teores mostrou-se como alternativa economicamente e ecologicamente

viável pois, quase sem exceção, este método não requer alto consumo de energia, não ocorre emissão de gases poluentes e formação de chuva ácida, devido à liberação óxidos de enxofre para atmosfera e não poluem efluentes aquáticos com resíduos de metais tóxicos, como acontece nos métodos convencionais [54].

Tendo em vista que a espectroscopia na região do infravermelho médio tem sido utilizada a mais de 40 anos para a caracterização de microrganismos baseando-se em seus diferentes espectros de infravermelho [62] e tem-se mostrado uma ferramenta útil para avaliar a diferença da composição química de bactérias em diferentes estágios de crescimento, em diferentes meios de cultura e na classificação em espécies e em subespécies [63], ela foi adotada neste trabalho como uma estratégia para avaliar as modificações sofridas por *A. ferrooxidans LR* quando submetida ao aumento da temperatura ótima de crescimento e à privação de fosfato. O uso desta abordagem foi encorajado pelo trabalho desenvolvido por Yu e Irudayaraj [64] onde mostraram que o citoplasma e o envelope celular bacteriano apresentam características espectroscópicas diferentes.

Desta forma, o objetivo desta aplicação da tese foi empregar a espectroscopia na região do infravermelho médio e métodos quimiométricos de classificação de amostras para investigar se e como o envelope celular de *A. ferrooxidans LR* [65] é afetado mediante os estresses propostos, além de comparar a eficiência dos próprios métodos de classificação SIMCA e SVM.

5.1 Experimental

Foi utilizada para os experimentos a linhagem *A. ferrooxidans LR* [65] isolada em efluente ácido de coluna de lixiviação de minério de urânio, em Lagoa Real, MG, Brasil. As bactérias foram cultivadas em Erlenmeyers de 250 mL em um shaker rotatório à 250 rotações por minuto em 100 mL de meio líquido contendo: 0,4 g/L de $K_2HPO_4.3H_2O$, 0,4 g/L de $MgSO_4.7H_2O$, 0,4 g/L de $(NH_4)_2SO_4$ e 33,4 g/L de $FeSO_4.7H_2O$ em pH 1,8 ajustado com ácido sulfúrico. As bactérias foram cultivadas sob condições controle (30ºC e presença de K_2HPO_4 no meio), sob

estresse térmico (40ºC) e em condições limitantes de fosfato (ausência de K_2HPO_4 no meio). As bactérias foram cultivadas até 50% da oxidação do ferro nos meios, monitoradas por titulação do íon ferroso com dicromato de potássio. Foram inoculadas 0,75 $x10^9$ bactérias por meio de cultura. As culturas obtidas foram filtradas em papel de filtro comum e a seguir, as células foram coletadas através de filtração em membrana Millipore (0,45 μM).

Para a realização deste experimento foram obtidas 47 amostras de massa celular, sendo 12 amostras cultivadas a 40ºC, 11 amostras cultivadas na privação de fosfato e 24 amostras da condição controle, sendo produzidas um conjunto de 12 juntamente com cada conjunto anterior. Para cada tratamento, a massa celular obtida foi congelada em nitrogênio líquido e liofilizada. Todo esse procedimento foi realizado pelo Centro de Biologia Molecular e Engenharia Genética (CBMEG) da Unicamp.

Para obtenção dos espectros da massa celular foi utilizado um espectrômetro de infravermelho ABB-Bomem MB Series com acessório de reflectância difusa e utilizado o Sulfato de Cálcio como branco. Os espectros foram obtidos em número de onda de 400 a 3800 cm^{-1} com 4 cm^{-1} de resolução e foram feitos 64 *scans* por amostra. Para tratamento dos dados e construção dos modelos quimiométricos foram utilizados o Matlab 7.0.1, PLS-Toolbox 4.02 e a rotina de SVM com otimização por GA proposta por Huang e Wang [36].

5.2 Avaliação de diferenças na estrutura celular de bactérias cultivadas a 30 e 40ºC.

Nesta primeira parte da aplicação, os métodos quimiométricos de classificação de amostras foram utilizados para avaliar diferenças celulares em *A. ferrooxidans* LR cultivada a 30 e 40˚C. A Figura 26 mostra os espectros obtidos com as réplicas experimentais.

Analisando a Figura 26 percebe-se que, visualmente, os espectros dos dois grupos de amostra (30 e 40˚C) não apresentaram nenhuma distinção aparente.

Para se tentar detectar diferenças entre esses grupos de amostras foram utilizados os algoritmos SIMCA e o SVM.

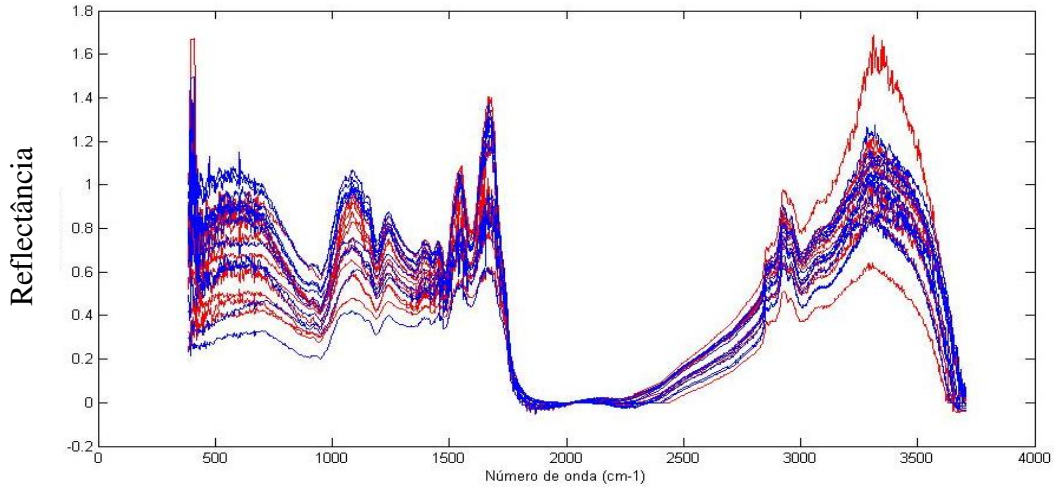


Figura 26. Espectros, com linha de base acertada, obtidos a partir de *A. ferrooxidans* cultivadas a 30˚C (em vermelho) e a 40˚C (em azul).

Para a construção dos modelos foram utilizadas as primeiras derivadas dos espectros e esses dados foram centrados na média, a fim de diminuir efeitos de matriz das amostras sobre o modelo. A Figura 27 mostra a primeira derivada dos espectros.

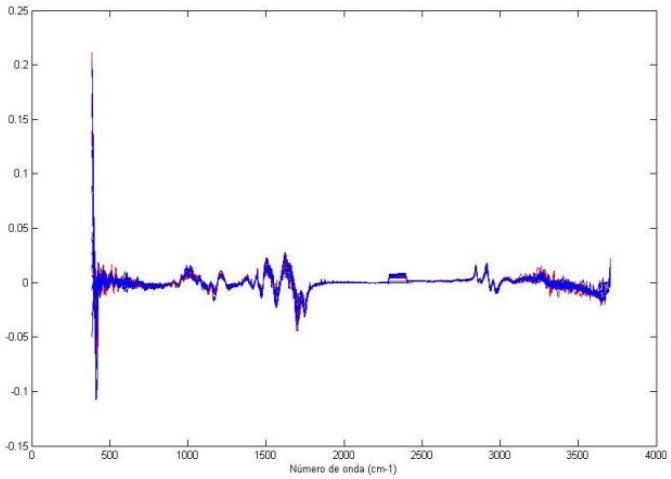


Figura 27. Primeira derivada dos espectros obtidos a partir de células secas de *A. ferrooxidans LR* cultivadas a 30˚C (em vermelho) e a 40˚C (em azul).

Assim, foram construídos modelos baseados na primeira derivada do espectro inteiro. Porém, não foram obtidos bons resultados para a separação dos dois grupos de bactérias. Então os espectros foram divididos em regiões e novamente foram utilizadas as primeiras derivadas centradas na média como dados de entrada no algoritmo. Modelos exploratórios foram preparados utilizando o iPCA e a região que apresentou melhores resultados foi a região situada entre 850 e 1275 cm^{-1} (Figura 28), que é a região referente ao *fingerprint* metabólico. Essa região detecta possíveis mudanças decorrentes da resposta a um determinado fator podendo elucidar alterações metabólicas, sem contudo detalhar vias bioquímicas [66].

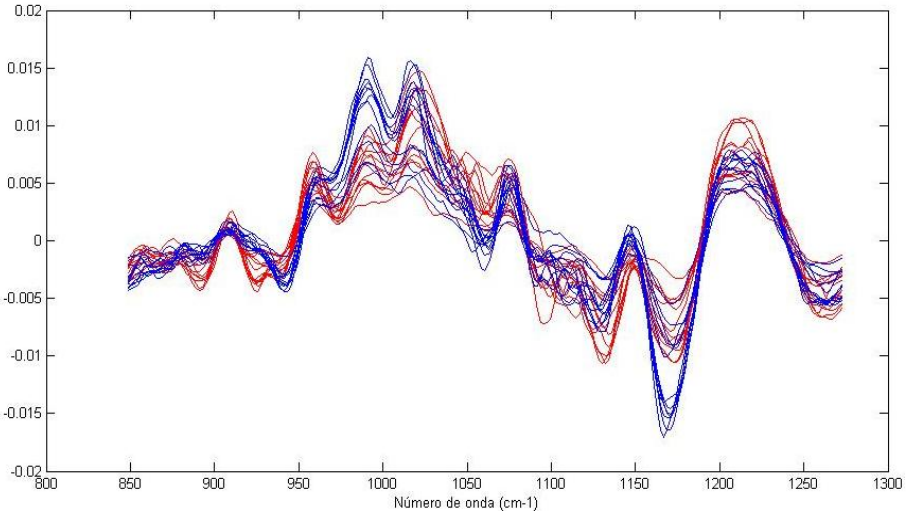


Figura 28. Primeira derivada da região dos espectros usada na construção dos modelos. Em vermelho espectros da *A. ferrooxidans LR* cultivada a 30˚C e em azul espectros das bactérias cultivadas a 40˚C.

Depois destes pré-tratamentos os dados foram separados em conjuntos de calibração e validação pelo algoritmo de Kennard-Stone, com 16 amostras no conjunto de calibração (com oito amostras referentes a cada condição de crescimento) e oito no conjunto de validação (com quatro amostras referentes a cada condição de crescimento).

5.2.1 Modelo SIMCA

Na construção do modelo SIMCA são ajustados modelos de PCA para cada uma das duas classes. Neste caso foram escolhidas 4 variáveis latentes tanto para a classe 1 (referente as bactérias cultivadas a 30ºC) que explicavam 92,17% da variância total, quanto para a classe 2 (referente as bactérias cultivadas a 40ºC) que explicavam 94,01% da variância total.

Na Figura 29 estão representadas as previsões feitas pelo SIMCA para as duas condições de crescimento, onde as amostras do conjunto de calibração para bactérias cultivadas à 30ºC estão em vermelho, amostras do conjunto de calibração para bactérias cultivadas à 40ºC estão em azul e amostras de validação dos mesmos conjuntos estão respectivamente em verde e rosa. As amostras que ficam na posição 1 são aquelas que o modelo previu como pertencentes a classe modelada. As mesmas informações aparecem na Tabela 6.

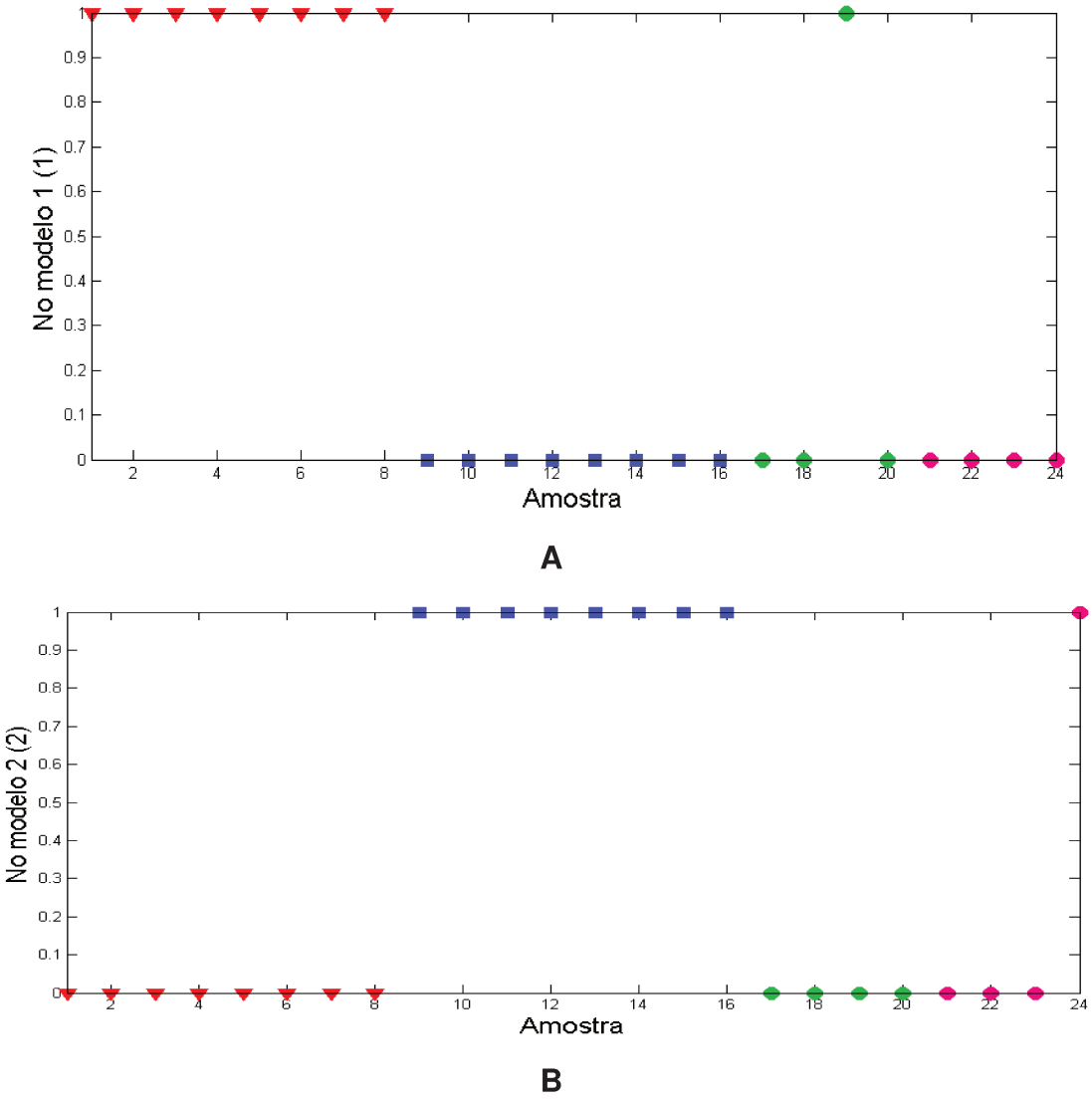


Figura 29. Previsão das classes das bactérias por SIMCA. Previsão para bactérias cultivadas a **A)** 30ºC e **B)** 40ºC.

Tabela 5. Legenda da Figura 29.

Classe	Tipo	Simbolo
30ºC	Calibração	
40ºC	Calibração	
30ºC	Validação	
40ºC	Validação	

Tabela 6. Previsão das classes das bactérias por SIMCA

Amostra	Temperatura de crescimento real (ºC)	Temperatura de crescimento prevista (ºC) A	Temperatura de crescimento prevista (ºC) B
Amostras de calibração			
1	30	30	-
2	30	30	-
3	30	30	-
4	30	30	-
5	30	30	-
6	30	30	-
7	30	30	-
8	30	30	-
9	40	-	40
10	40	-	40
11	40	-	40
12	40	-	40
13	40	-	40
14	40	-	40
15	40	-	40
16	40	-	40
Amostras de validação			
17	30	-	-
18	30	-	-
19	30	30	-
20	30	-	-
21	40	-	-
22	40	-	-
23	40	-	-
24	40	-	40

Podemos observar que nas previsões das classes das amostras dos grupos de calibração o SIMCA foi bastante eficiente, não havendo nem uma amostra classificada erroneamente; entretanto, a previsão das classes das amostras de validação foi bastante falha. Em ambas não houveram falsos positivos, porém, das 4 amostras que deveriam ter sido previstas como pertencentes a cada classe, apenas uma de cada foi prevista corretamente.

5.2.2 Modelo SVM

Os parâmetros do SVM foram otimizados por algoritmo genético utilizando 10 gerações, sendo obtidos como parâmetros ótimos γ=1024 e σ^2=32. Com esse modelo a validação cruzada teve 100% de acerto das classes das amostras, ou seja, a previsão das classes feita pelo algoritmo foi correta para todas as amostras de calibração do modelo.

Utilizando este mesmo modelo para prever as amostras de validação houve apenas um erro dentre as 8 amostras, como pode ser visto na Tabela 7.

Tabela 7. Temperaturas de crescimento reais e previstas pelo SVM nas amostras de validação.

Amostra	Temperatura de crescimento real (ºC)	Temperatura prevista (SVM) (ºC)
1	30	30
2	30	40
3	30	30
4	30	30
5	40	40
6	40	40
7	40	40
8	40	40

O SVM foi capaz de fazer uma ótima separação entre os dois grupos de amostras, o que pode ser corroborado na Tabela 7. Na Figura 30 ficam evidenciadas distinções entre os grupos de amostras, principalmente nos números de onda próximos a 990 e 1170 cm^{-1}.

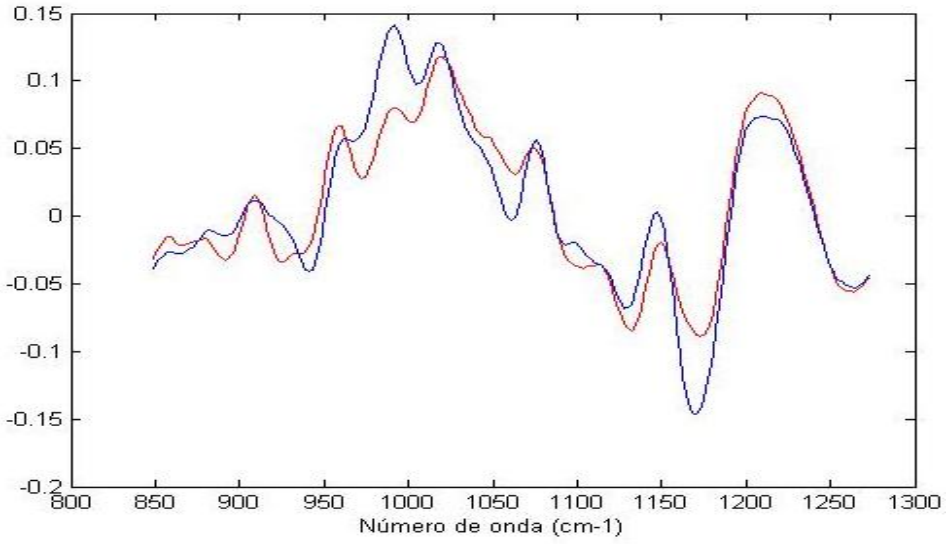


Figura 30. Média da primeira derivada da região do infravermelho usada na construção dos modelos. Em vermelho espectros da *A. ferooxidans LR* cultivada a 30˚C e em azul espectros das bactérias cultivadas a 40˚C.

Essa é a região correspondente aos polissacarídeos, sendo que bandas em 1160 ± 30 cm^{-1} são dominadas por ligações glicosídicas do tipo C-O-C [67], indicando possível aumento na produção de polissacarídeos por *A. ferrooxidans* cultivada em temperatura superior a ideal.

A parede celular e a membrana externa são as regiões com maior concentração de polissacarídeos em bactérias (principalmente fazendo parte do LPS – lipopolissacarídeo), indicando que essas estruturas podem ter sido afetadas pelo aumento da temperatura de crescimento.

Além da já citada indução de proteínas de choque térmico, bactérias respondem a situações estressantes com o ajuste da composição de lipídeos em suas membranas [68]. Esse ajuste da composição de lipídeos pode ser detectado, por infravermelho, através da deformação da banda que ocorre em torno de 1650

cm^{-1} (característico de C=C), devido a modificações na instauração dos lipídeos ou através de modificações na região dos ácidos graxos, em torno de 3000 cm^{-1}. Neste último caso, são observadas mudanças na posição da banda correspondente ao CH_2 [69]. Curiosamente, não foi detectada nenhuma modificação significativa nas regiões citadas acima, o que nos faz propor que em *A. ferrooxidans LR* o ajuste na composição de polissacarídeos possa ser crucial para a adaptação em temperaturas elevadas.

5.3 Avaliação de diferenças na estrutura celular de bactérias cultivadas na presença e ausência de Fósforo.

Nesta aplicação os métodos quimiométricos de classificação de amostras foram utilizados para avaliar diferenças celulares em *A. ferrooxidans LR* cultivada na presença e ausência de fosfato. A Figura 31 mostra os espectros obtidos com as réplicas experimentais.

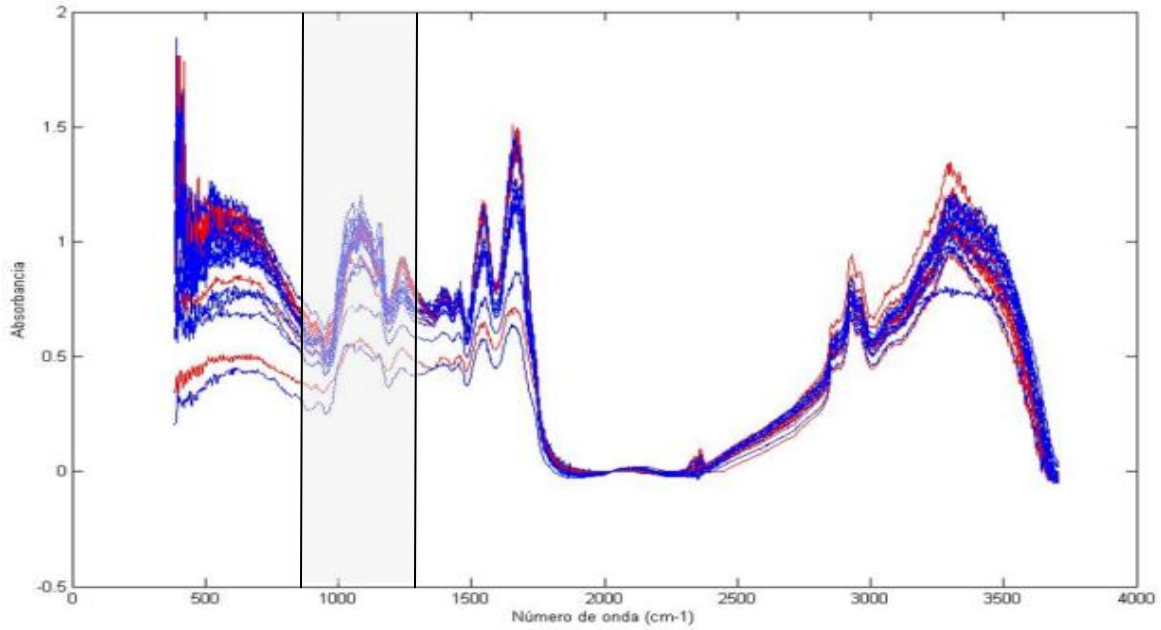


Figura 31. Espectros, com linha de base ajustada, obtidos a partir de células secas de *A. ferrooxidans* LR cultivada em presença (em vermelho) e ausência de fosfato (em azul). Em destaque a região utilizada na construção dos modelos.

Assim, como na aplicação anterior, os espectros não apresentaram diferenças aparentes entre as leituras de amostras com diferentes condições de crescimento.

Para a construção dos modelos foram utilizadas as primeiras derivadas dos espectros e esses dados foram centrados na média, a fim de diminuir efeitos do espalhamento de radiação sobre o modelo.

Inicialmente, foram construídos modelos baseados na primeira derivada do espectro inteiro e como no caso anterior não foram obtidos bons resultados para a separação dos dois grupos, sendo necessária a divisão dos espectros em regiões. Modelos exploratórios foram preparados utilizando o iPCA e a região que apresentou melhores resultados foi novamente a região referente ao "*fingerprint*" metabólico, situada entre 850 e 1275 cm^{-1}. A primeira derivada dessa região está presente na Figura 32.

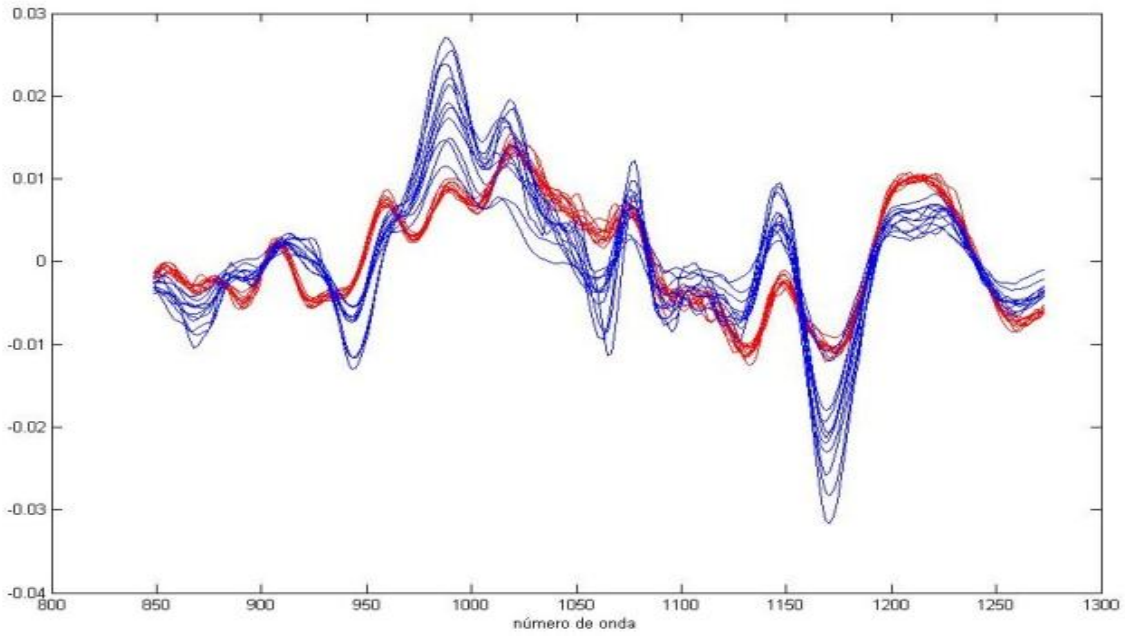


Figura 32. Primeira derivada da região do infravermelho usada na construção dos melhores modelos de previsão*,* Em vermelho espectros da *A. ferooxidans LR* cultivada na presença e em azul espectros das bactérias cultivadas na ausência de fosfato.

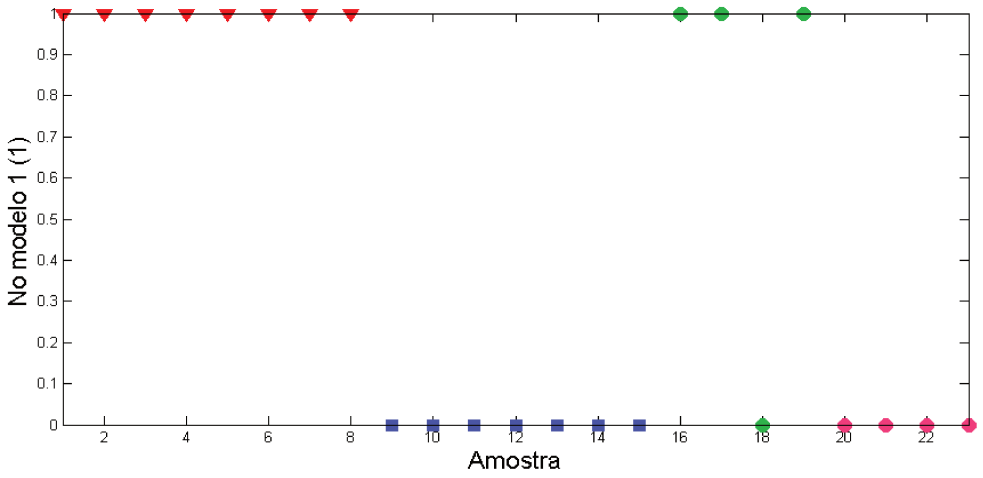
Podemos observar nesta figura algumas diferenças entre os grupos de amostras, como a região do espectro são a correspondente ao estiramento P=O de fosfodiésteres PO^{2-} (compreendida entre 1220 e 1260 cm^{-1}). Também se podem notar diferenças em bandas como a 1160 ± 30 cm^{-1} que é atribuída por ligações glicosídicas do tipo C-O-C [67].

Para validação dos modelos de classificação, depois destes pré-tratamentos os dados foram separados em conjuntos de calibração e validação pelo algoritmo de Kennard-Stone, com 15 amostras no primeiro conjunto (com 8 amostras referentes as bactérias cultivadas na presença de fosfato e 7 amostras referentes as bactérias cultivadas na ausência de fosfato) e 8 no segundo (com 4 amostras referentes a cada condição de crescimento).

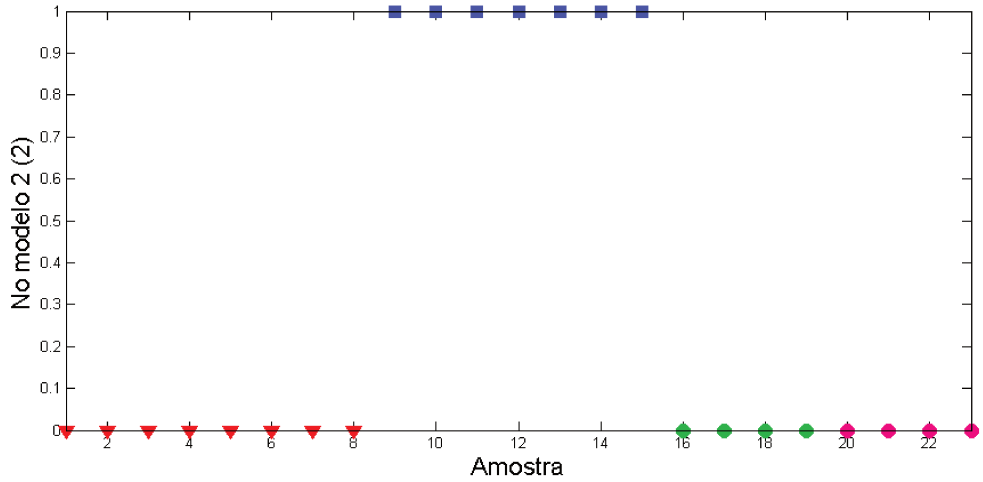
5.3.1 Modelo SIMCA

Na construção do modelo SIMCA foram ajustados os modelos de PCA para cada uma das duas classes, utilizando 4 variáveis latentes tanto para a classe 1 (referente as bactérias cultivadas na presença de fosfato) onde explicavam 93,24% da variância total, quanto para a classe 2 (referente as bactérias cultivadas na ausência de fosfato) onde explicavam 94,01% da variância total.

Na Figura 33 estão representadas as previsões feitas pelo SIMCA para as duas classes de bactérias, onde as amostras do conjunto de calibração para bactérias da classe 1 estão em vermelho, amostras do conjunto de calibração para classe 2 estão em azul e as respectivas amostras de validação estão em verde e rosa. As amostras que ficam na posição 1 são aquelas que o modelo previu como pertencentes a classe modelada. As mesmas informações estão presentes na Tabela 9.



A



B

Figura 33. Previsão de classes por SIMCA para bactérias cultivadas **A)** na presença de fosfato e **B)** na ausência de fosfato.

Tabela 8. Legenda da Figura 33.

Classe	Tipo	Simbolo
Presença de fosfato	Calibração	
Ausência de fosfato	Calibração	
Presença de fosfato	Validação	
Ausência de fosfato	Validação	

Tabela 9. Previsão das classes das bactérias por SIMCA.

Amostra	Condição real	Condição prevista A	Condição prevista B
Amostras de calibração			
1	Presença	Presença	-
2	Presença	Presença	-
3	Presença	Presença	-
4	Presença	Presença	-
5	Presença	Presença	-
6	Presença	Presença	-
7	Presença	Presença	-
8	Presença	Presença	-
9	Ausência	-	Ausência
10	Ausência	-	Ausência
11	Ausência	-	Ausência
12	Ausência	-	Ausência
13	Ausência	-	Ausência
14	Ausência	-	Ausência
15	Ausência	-	Ausência
Amostras de validação			
16	Presença	Presença	-
17	Presença	Presença	-
18	Presença	-	-
19	Presença	Presença	-
20	Ausência	-	-
21	Ausência	-	-
22	Ausência	-	-
23	Ausência	-	-

Como no modelo anterior, nas previsões dos grupos de calibração das duas classes o SIMCA foi bastante preciso e não errou a classificação de nenhuma

amostra. Quanto às amostras de validação, o modelo previu bem as amostras cultivadas na presença de fosfato, não classificando corretamente apenas uma amostra, entretanto não foi capaz de prever nenhuma das 4 amostras cultivadas na ausência de fosfato. Não houve nenhum falso positivo em nenhum caso.

5.3.2 Modelo SVM

Os parâmetros do SVM foram otimizados por algoritmo genético, utilizando 10 gerações. Os parâmetros obtidos foram γ=1024 e σ^2=0,2500. Com esse modelo a validação cruzada teve 100% de acerto das classes das amostras, ou seja, a previsão de todas as amostras de calibração foi correta.

O modelo preparado foi utilizado para prever as amostras de validação, e não houve nenhum erro de classificação, como pode ser visto na Tabela 10.

Tabela 10. Condições de crescimento reais e previstas pelo SVM.

Amostra	Condição de crescimento real	Condição prevista (SVM)
1	Presença de Fosfato	Presença de Fosfato
2	Presença de Fosfato	Presença de Fosfato
3	Presença de Fosfato	Presença de Fosfato
4	Presença de Fosfato	Presença de Fosfato
5	Ausência de Fosfato	Ausência de Fosfato
6	Ausência de Fosfato	Ausência de Fosfato
7	Ausência de Fosfato	Ausência de Fosfato
8	Ausência de Fosfato	Ausência de Fosfato

O SVM foi capaz de fazer uma ótima classificação entre os dois grupos de amostras, com nenhum erro de previsão, o que pode ser evidenciado na Tabela 10.

5.4 Conclusões

Em ambas as aplicações de infravermelho médio e quimiometria para classificação de bactérias, em relação a sua condição de crescimento, o SVM mostrou-se uma ferramenta mais eficiente, obtendo resultados de previsão da classe das amostras de validação mais coerentes com as classes reais das mesmas do que o algoritmo padrão na área de quimiometria que é o SIMCA.

Além disso, a diferença de classes detectada mostrou que há, de fato, uma mudança estrutural na *A. ferrooxidans LR* causada por estresses de temperatura e privação de fosfato. A região do espectro que originou os melhores modelos, ou seja, maiores separações de classes, indica que essas mudanças são causadas por alterações nos carboidratos, fosfolipideos e fosfoproteínas, que devem funcionar como sistema de defesa da bactéria a essas situações.

Capítulo VI

6. Diferenciação entre nódulos hepáticos por microespectroscopia no MIR

Hepatocarcinoma (HCC) [70] é o tipo mais comum de câncer do fígado, representando mais de 90% dos casos. A relação entre o HCC e a cirrose está bem estabelecida, especialmente nos casos secundários às infecções pelos vírus B e C da hepatite e ao consumo de bebidas alcoólicas [71].

Nódulos hepáticos podem ser benignos ou sofrer transformação para HCC. Há uma seqüência de nódulos hepáticos, regenerativos ou cirróticos, displásicos de baixo grau e displásicos de alto grau, que precedem o HCC. As lesões são definidas de forma consistente apenas por meio de análise histológica. Com o conhecimento que a hepatocarcinogênese segue determinada sequência de eventos e, visando alternativas terapêuticas mais precoces e eficientes, na atualidade tem-se buscado cada vez mais a identificação de lesões hepáticas pré-neoplásicas ou com algum potencial de se transformarem em HCC, principalmente através de métodos de imagem. Porém, apesar de muito valorizadas e estudadas, do ponto de vista histológico há ainda grande confusão diagnóstica, muito em parte devido à falta de critérios morfológicos unânimes, além de grande confusão na sua nomenclatura, com várias classificações propostas [72].

Desde 1995, a classificação mais utilizada para estas lesões é a proposta pela *International Working Party* [73], que as classifica em: MacroNódulo Regenerativo (MNR), Nódulo displásico de baixo grau (NDBG) Nódulo Displásico de Alto Grau (NDAG) e Carcinoma Hepatocelular. O diagnóstico diferencial entre MNR e NDBG, em alguns casos, é extremamente difícil, quando não impossível. Porém, como parece não haver significado prático para esta diferenciação, alguns autores propõem que conjuntamente sejam designadas como “lesão hepatocelular de baixo grau”. O desenvolvimento e evolução do HCC está apresentado na Figura 34 [72].

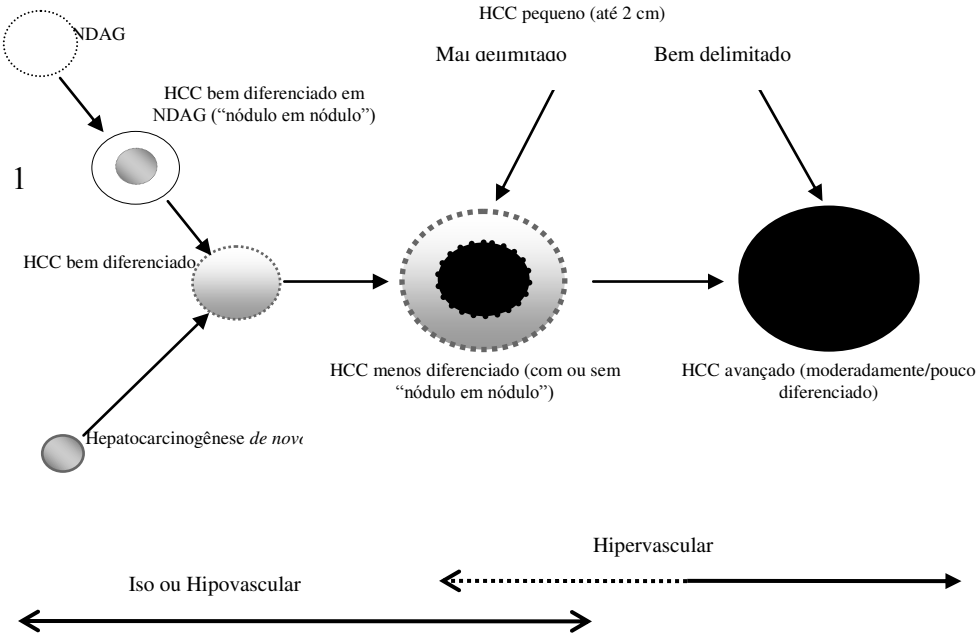


Figura 34. Desenvolvimento e evolução do HCC.

O fato de macronódulos serem bem mais frequentes em fígados cirróticos com HCC demonstra forte relação entre estes. Além disso, alguns estudos têm demonstrado o aparecimento de HCC no interior de macronódulo. De acordo com a literatura [72], ocorreria um processo clonal no interior destas lesões, em que clones de células malignas originando-se de hepatócitos que sofreram mutações, substituiriam gradativamente todo o nódulo, transformando-o em HCC bem diferenciado. O diagnóstico diferencial entre NDAG e HCC pequeno é também muito difícil. A demonstração, no entanto, que muitos HCC se originam no interior de NDAG faz com que vários cirurgiões optem por tratá-los igualmente

Vale comentar que o diagnóstico diferencial de nódulos benignos e do HCC se faz necessário para que os pacientes possam receber a terapêutica adequada a cada caso. Ainda, o diagnóstico precoce do HCC confere melhor prognóstico a seus portadores. Entretanto, são comuns as dificuldades encontradas para a diferenciação de nódulos hepáticos benignos e malignos (HCC) por meio da análise histológica convencional, o que justifica a busca por outros métodos que contribuam para o diagnóstico correto das lesões distintas [74,75].

Têm sido desenvolvidos métodos utilizando espectroscopia na região do NIR para diagnóstico de câncer de próstata [76], mama [77] e cólon [78], assim,

nesta quarta aplicação, utilizou-se a microespectroscopia na região do infravermelho próximo [79] em conjunto com métodos quimiométricos de classificação para diferenciação de nódulos cirróticos e HCC. Também foi realizada uma comparação no poder de classificação entre o SIMCA e o SVM.

A microespectroscopia no infravermelho refere-se ao acoplamento de um espectrômetro a um microscópio. Os espectros podem ser obtidos de apenas um ponto da amostra ou para se obter uma imagem global da amostra. A microscopia no infravermelho é uma técnica capaz de diferenciar características de amostras em nível microscópico, podendo revelar distribuições e constituintes da mesma [79].

No microespectômetro a luz da fonte é focalizada sobre a amostra utilizando um condensador e a luz transmitida ou refletida pela amostra é coletada pela objetiva, formando uma imagem ampliada da amostra. Essa imagem é então levada à um detector apropriado. No geral, a função e componentes encontrados no microespectômetro não diferem de um microscópio convencional; as únicas exceções são que o microespectômetro IR: (1) emprega radiação infravermelha do interferômetro como sua fonte, (2) utiliza lentes refletindo, (3) utiliza uma abertura para o plano da imagem primária, para definição da amostra e (4) utiliza um detector sensível ao infravermelho [80].

Nos primeiros anos de microespectroscopia no infravermelho a técnica foi empregada principalmente para a identificação de contaminantes particulados. Foi logo reconhecido, no entanto, que o método poderia ser empregado para obter informações químicas de áreas localizadas de uma amostra muito maior para obter a distribuição das espécies químicas na mesma. Naquela época, microscópios IR foram equipados com estágios x-y controlada por computador que permitia obtenção de mapas químicos [80].

6.1 Experimental

Foram analisados fragmentos de fígado humano, cirróticos ou HCC, incluídos em blocos de parafina do arquivo do Gastrocentro da Universidade

Estadual de Campinas. Os cortes histológicos de 5 μm de cada caso foram montados em lâminas de vidro. Os fragmentos foram analisados em um espectrômetro SPOTLIGHT 400N Perkin Elmer utilizando os parâmetros: modo imagem, transmitância, 64 *scans* por pixel, resolução 4 cm^{-1}, tamanho de pixel de 25 μm^2, faixa espectral de 2000 a 6000 cm^{-1}, área de 100 x 100 μm, utilizando a própria lâmina de vidro como branco.

Para a construção dos modelos foram utilizados espectros de 83 nódulos de HCC e 41 de nódulos cirróticos de diferentes graus, classificados através de análise histológica por especialistas do Gastrocentro da Universidade Estadual de Campinas. Para tratamento dos dados e construção dos modelos quimiométricos foram utilizados o Matlab 7.0.1, PLS-Toolbox 4.02 e a rotina de SVM com otimização por GA proposta por Huang e Wang [36].

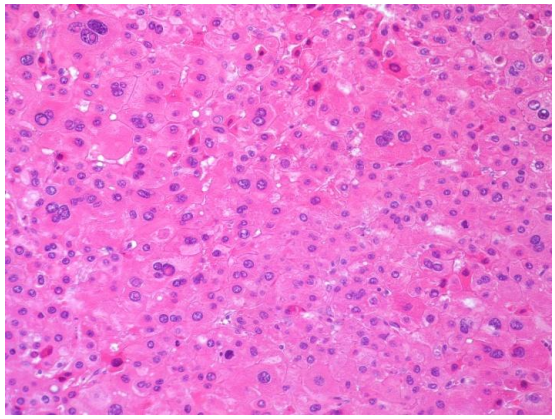


Figura 35. Exemplo de nódulo de HCC visto no microscópio acoplado ao equipamento de NIR.

6.2 Resultados e discussões

Os espectros utilizados na construção dos modelos quimiométricos foram a média de todos os espectros obtidos em cada imagem. Com isso tentava-se deixar os espectros das diversas amostras mais homogêneos entre si, diminuindo

possíveis alterações espectrais devido a diferenças estruturais presentes no próprio tecido.

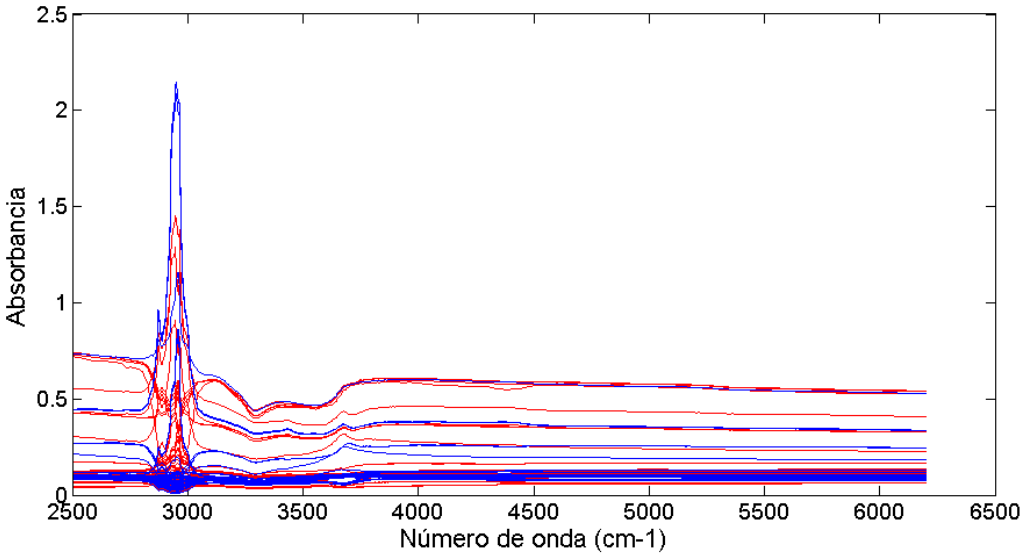


Figura 36. Espectros médios das amostras de nódulos cirróticos (em vermelho) e cancerosos (em azul).

Como é possível observar na Figura 36 há pouca informação a partir de 4000 cm^{-1}, por este motivo o espectro foi cortado entre 2500 e 4000 cm^{-1}, região pertencente ao infravermelho médio, antes de ser utilizado. A vantagem adicional de usar esta região do espectro é que sua absorção pelo vidro das lâminas é muito fraca ou nula.

Pode ser visto que os espectros tiveram bastante problema de linha de base. A fim de minimizar esse problema de espalhamento de radiação foi utilizado como pré-processamento a segunda derivada dos espectros. Também foi importante auto-escalar e normalizar as amostras para eliminar diferenças de matriz entre as amostras. Entretanto, modelos feitos apenas com esses pré-processamentos não foram capazes de separar as classes de amostras de forma eficaz, assim alguns outros pré-processamentos foram testados a fim de aumentar a exatidão da separação das classes.

O pré-processamento utilizado nos modelos que apresentaram melhores resultados, ou seja, uma classificação mais correta das amostras, foi a Correção

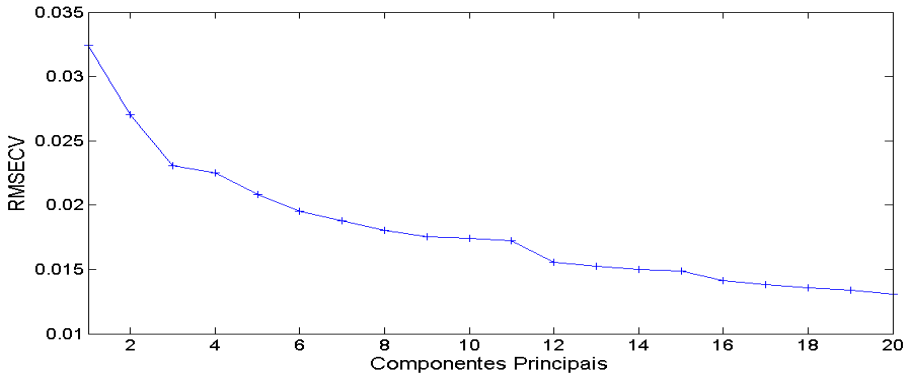
Ortogonal de Sinal (OSC, do nome em inglês *Orthogonal Signal Correction*) [20]. Este algoritmo é utilizado para eliminar informação desnecessária dos espectros. Em amostras complexas, como por exemplo, de origem biológica, este é o pré-processamento mais utilizado. Neste procedimento a matriz **X** é corrigida pela subtração da variação que é ortogonal ao vetor de calibração **y**. O **y** neste caso era um vetor contendo a classe correspondente a cada amostra.

Após passarem pelos pré-processamentos, os espectros das amostras foram separados aleatoriamente em um conjunto de calibração, com 99 amostras, e um conjunto de validação, com 25, sendo 16 de HCC e 9 de nódulos cirróticos.

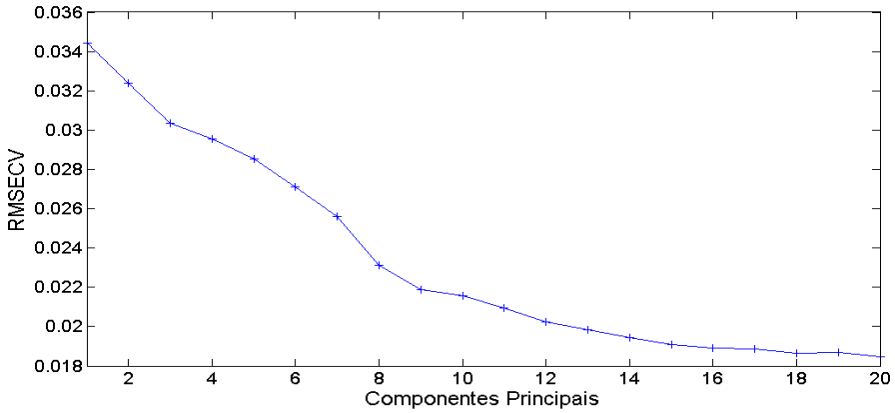
6.2.1 Modelo SIMCA

No SIMCA um modelo PCA é ajustado separadamente para cada classe. Para isso deve-se escolher o número de componentes principais de cada PCA utilizando a validação cruzada. A Figura 37 mostra os gráficos da raiz do erro médio quadrático de validação cruzada (RMSECV) por Componentes Principais para as duas classes de amostras.

Através da interpretação das Figuras 37A e B foram escolhidas 12 componentes principais para ambas as classes. Apesar de ser o mais indicado através da observação dos gráficos este é um número bastante alto de componentes principais, o que poderia levar o modelo a um sobreajuste.



A



B

Figura 37. RMSECV para **A)** nódulos cancerosos e **B)** nódulos cirróticos.

Com o modelo SIMCA ajustado foram construídos gráficos mostrando a previsão de cada classe. Na Figura 38 as amostras de calibração de carcinoma e nódulo cirrótico estão respectivamente em vermelho e azul, enquanto os conjuntos de validação estão em verde e rosa (Tabela 11). As amostras consideradas como pertencentes à classe modelada ficam na posição 1 do gráfico, assim era esperado que em A (previsão de HCC) ficassem na posição 1 apenas amostras vermelhas e verdes, enquanto em B (previsão de nódulos cirróticos) ficassem na posição 1 apenas as amostras azuis e rosas.

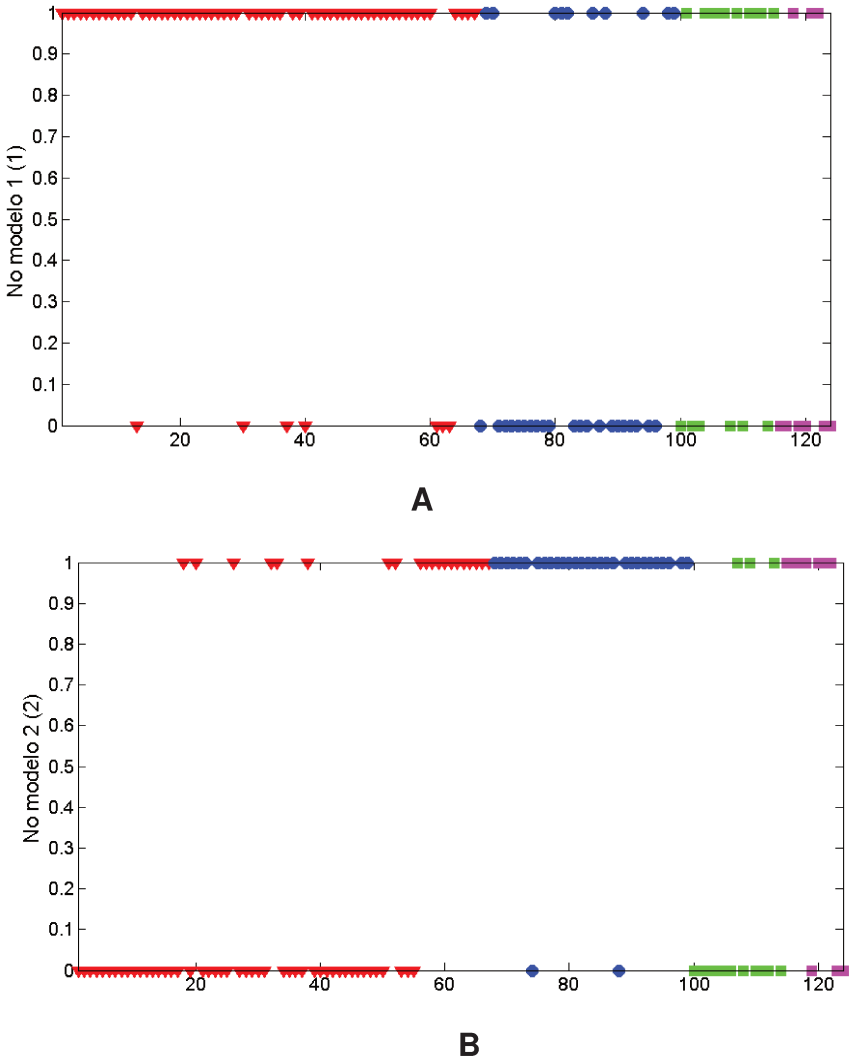


Figura 38. Previsão das amostras de corte histológico de fígado separadas em conjuntos de calibração e validação. Onde, em **A)** está a previsão das amostras de nódulos cancerosos e em **B)** cirróticos.

Tabela 11. Legenda da Figura 38.

Classe	Tipo	Simbolo
HCC	Calibração	
Nódulo cirrótico	Calibração	
HCC	Validação	
Nódulo cirrótico	Validação	

Como pode ser visto, a previsão se mostrou bastante falha, principalmente nos conjuntos de classificação. Quanto as amostras de validação para carcinoma, 6 das 16 amostras não foram previstas, além de 3 falsos positivos, enquanto na previsão para nódulos cirróticos houveram três amostras que não foram previstas, além de 3 falsos positivos.

A previsão das amostras de validação também está especificada na Tabela 12.

6.2.2 Modelo SVM

Os parâmetros foram otimizados por algoritmo genético, utilizando 15 gerações. Os parâmetros obtidos através do algoritmo genético foram γ=32768 e σ^2=0,0625. Com esse modelo a validação cruzada, ou seja, previsão das próprias amostras do conjunto de calibração, teve 81,82% de acerto das classes das amostras.

Utilizando esse modelo para prever as amostras de validação foram obtidos 76% de acerto geral. O modelo previu corretamente 13 das 16 amostras de validação de câncer e 6 das 9 das amostras de validação de nódulo cirrótico, como pode ser visto na Tabela 12. A mesma tabela também traz os resultados de previsão do modelo SIMCA, a fim de comparação.

Tabela 12. Classes reais e previstas por SVM das amostras de corte histológico de fígado.*

Amostra	***Classe Real***	***Classe Prevista SVM***	***Classe Prevista SIMCA***	***Classe Prevista SIMCA***
1	HCC	HCC	-	-
2	HCC	HCC	HCC	-
3	HCC	HCC	-	-
4	HCC	HCC	-	-
5	HCC	HCC	HCC	-
6	HCC	HCC	HCC	-
7	HCC	HCC	HCC	-
8	HCC	HCC	HCC	-
9	HCC	HCC	-	Nod
10	HCC	Nod	HCC	-
11	HCC	Nod	HCC	Nod
12	HCC	HCC	HCC	-
13	HCC	Nod	-	-
14	HCC	HCC	HCC	-
15	HCC	HCC	-	Nod
16	HCC	HCC	HCC	-
17	Nod	HCC	-	Nod
18	Nod	Nod	-	Nod
19	Nod	HCC	HCC	Nod
20	Nod	Nod	-	-
21	Nod	Nod	-	Nod
22	Nod	HCC	HCC	Nod
23	Nod	Nod	HCC	Nod
24	Nod	Nod	-	-
25	Nod	Nod	-	-

* onde Nod eram as amostras classificadas como nódulos cirróticos.

6.3 Conclusões

Através dos resultados pode-se notar que, devido a complexidade das amostras, o SVM se mostrou mais eficiente do que o SIMCA, especialmente na classificação das amostras de nódulos HCC, onde o SIMCA apresentou o dobro de erros do que o SVM nas amostras de validação.

O SVM é bastante dependente do número de amostras, com um conjunto de amostras maior a tendência é que houvesse um aumento da taxa de acertos na previsão da classe dos nódulos com este algoritmo. Entretanto, esse modelo já foi bastante satisfatório, uma vez que em um teste clínico são utilizadas em média 5 amostras de um mesmo nódulo antes de dar um diagnóstico.

A microespectroscopia na região do infravermelho aliada ao SVM parece ser uma alternativa interessante na diferenciação dos nódulos hepáticos, sendo mais rápida do que a análise histológica convencional e principalmente por não necessitar de um profissional altamente treinado para realizá-la.

Conclusões Gerais

7. Conclusões

Nesta tese foram realizadas quatro aplicações do algoritmo *Support Vector Machines*, sendo duas para calibração multivariada e duas para classificação de amostras, para realizar uma comparação de performance deste com a de outros algoritmos mais utilizados para estes fins.

Os estudos onde o SVM foi utilizado para a calibração multivariada foram a determinação de Nitrogênio e Carbono em solo e a determinação da concentração de um determinado mineral em um polímero.

Na primeira utilização do SVM, ele se mostrou uma ferramenta mais indicada para a determinação da concentração do Carbono e Nitrogênio quando estes elementos estão em solos sem que se saiba se há ou não a presença do mineral gipsita.

Nos modelos construídos com as amostras de solo sendo separadas em grupos com e sem o mineral, o PLS foi capaz de prever a concentração tão bem quanto o SVM, não havendo indícios estatísticos de diferença de desempenho entre os algoritmos. Por outro lado, nos modelos construídos com amostras de solos com e sem gipsita juntas, as concentrações desses elementos foram obtidas com erros consideravelmente menores do que a previsão feita pelo método padrão de calibração multivariada, o PLS. Ou seja, o SVM parece ser uma boa alternativa para construção de modelos de calibração quando há mais de um tipo de amostra presente nos conjuntos de calibração e validação.

Na determinação da concentração de um determinado mineral em polímero o PLS conseguiu boas previsões, contudo o modelo SVM conseguiu chegar a um modelo com um erro médio quadrático de previsão com um valor de praticamente metade do primeiro algoritmo. Observando o teste F feito entre esses RMSEPs e o gráfico de erros absolutos das previsões pudemos concluir que o SVM produziu previsões significantemente melhores que o PLS.

Nos estudos de classificação de amostras biológicas, o SVM teve um desempenho bastante superior ao método padrão SIMCA. A complexidade das amostras conferia uma característica não normalmente encontrada em problemas

deste tipo e devido a isso a correta classificação das mesmas se torna bastante comprometida com a utilização de métodos mais tradicionais de reconhecimento de padrões, como SIMCA.

Na classificação das bactérias quanto às condições de crescimento (temperaturas 30 ou 40°C e na presença ou ausência de fosfato) o SIMCA não foi capaz de classificar corretamente a grande maioria das amostras de validação enquanto o SVM produziu apenas uma previsão errada para as quatro diferentes condições de crescimento das bactérias nestes conjuntos.

Através desta classificação entre os grupos de bactérias foi evidenciado que a *A. ferrooxidans LR* se modifica, e principalmente em componentes da parede celular, quando submetida a essas mudanças nas condições ambientais. Essa adaptação é uma informação importante, pois mostra que a bactéria sobrevive bem às intempéries do processo de biolixiviação e indica pontos que futuramente podem ser explorados para melhorar ainda mais o seu potencial biotecnológico.

Quanto à diferenciação de nódulos cirróticos e de HCC, houve uma considerável diferença na previsão do HCC com o SVM produzindo o dobro de acertos em relação ao SIMCA.

A taxa de previsões corretas feitas pelo SVM permite que a utilização deste algoritmo em dados de microscopia NIR de laminas de fígado possa ser proposta como método alternativo de diagnóstico de diferenciação dos nódulos hepáticos entre cirróticos e carcinogênicos, isso traz como vantagens o desenvolvimento de uma metodologia mais rápida do que a análise histológica convencional e principalmente por não necessitar de um profissional altamente treinado para realizá-la.

Nos 4 estudos feitos nessa tese o SVM acabou por ter uma performance melhor do que os algoritmos aos quais foi comparado, tanto para os casos onde foi utilizado para desenvolver modelos de calibração multivariada quanto para os casos onde foi utilizado para classificação de amostras, indicando ser uma alternativa promissora para quando algoritmos tradicionais não produzam resultados satisfatórios.

Bibliografia

8. Bibliografia

[1] Skoog, D.; Princípios de Análise Instrumental, Bookman, Porto Alegre, 2002.

[2] Vandegniste, B.G.M.; Massat, D.L.; Buydens, L.M.C.; Jong, S.; Lewi, P.J.; Verbeque, J.S.; Handbook of chemometrics and qualimetrics: part B., Elsevier, Amsterdam, 1998.

[3] Miller, J.N.; Miller, J.C.,;Statistics and chemometrics for analytical chemistry, Chinchester: Prentice Hall, 2000.

[4] Coates, J.; Spectroscopy 14 (1999) 20.

[5] Brereton, R.; Analyst (2000) 2125.

[6] Otto, M.; Chemometrics, wiley -VCH, New York, 1999.

[7] Brereton, R.; Analyst 112 (1987) 1635.

[8] Geladi, P.; Kowalski, B.R.; Anal. Chim. Acta 185 (1986) 17.

[9] Martens, N.; Naes, T.; Multivariate calibration, John Wiley & Sons, New York, 1989.

[10] Massart, B.; Vandegiste, S.; Deming, S.N; Chemometrics: a text book, Elsevier, Amsterdam, 1988.

[11] Mellinger, M.; Chemom. Intell. Lab. Syst. 2 (1987) 29.

[12] Wold, S.; Esbensen, K.; Geladi, P.; Chemom. Intell. Lab. Syst. 2 (1987).

[13] Thomas, E.; Haaland, D.; Anal. Chem. 62 (1990) 1091.

[14] Draper, N.R.; Applied Regression Analysis, John Wiley & Sons, New York, 1981.

[15] Geladi, P.; Martens, H.; Appl. Spectrosc. 39 (1985) 491.

[16] Isaksson, T.; Appl. Spectrosc. 42 (1988).

[17] Barnes, R. J. ; Lister, S. J. ; Appl. Spectrosc. 43 (1989) 772.

[18] Bracewell, R.; The Fast Fourier Transform and its aplication, McGraw-Hill, New York, 1965.

[19] Bouveresse, E.; Casolino, C.; Massart, D. L.; Appl. Spectrosc. 52 (1998) 604.

[20] Wold, S.; Antti, H.; Lindgren, F.; Ohman, J.; Chemom. Intell. Lab. Syst. 44 (1998) 175.

[21] Gavaghan, C.; Wilson, I.D.; Nicholson, J.; FEBS Letters 550 (2002) 191.

[22] Abrahamsson, C. ; Chemom. Intell. Lab. Syst. 69 (2003) 3.

[23] Costa Filho, P.A.; Poppi, R.J.; Quim. Nova 22 (1999) 405.

[24] Zupan, J.; Gasteiger, J.; Neural Networks for Chemistry: an introduction, Weinheim: VCH, 1993.

[25] Wise, B. M.; Bro, R; Shaver, J. M.; Windig, W.; Koch, R. S.; Eingenvector research Inc., 2005.

[26] Poppi, R.; Braga, J. W. B.; Quim. Nova 27 (2007) 1004.

[27] de Sena, M. M.; Trevisan, M. G.; Poppi, R. J.; Talanta 68 (2006) 1707.

[28] Ferrão, M.F.; Mello, C.; Borin, A.; Maretto, D.A.; Poppi, R.J.; Quim. Nova 30 (2007) 852.

[29] Haykin, S.; Redes Neurais - Princípios e prática, Bookman, Porto Alegre, 2001.

[30] Li, H.; Chemom. Intell. Lab. Syst. 95 (2009) 188.

[31] Borin, A.; Ferrão, M.F., Mello, C. Maretto, D.A., Poppi, R.J., Anal. Chem. Acta 579 (2006) 25.

[32] Thissen, U.; Üstün, B.; Melssen, W.J.; Buydens, L.M.C.; Anal. Chem. 76 (2004) 3099.

[33] Üstün, B.; Melssen, W.; Buydens, L.; Chemom. Intell. Lab. Syst. 81 (2006) 26.

[34] Cogdill, R.; Dardenne, P.; J.Near Infrared Spectroc. 12 (2004) 93.

[35] Cortes, C.;Vapnik, V.; Mach. Learn. 20 (1995) 273.

[36] Huang, C.; Wang, C.; Expert Syst Appl 31 (2006) 231.

[37] Fearnside, P.; Barbosa, R.; Forest Ecol Manag 108 (1998) 147.

[38] He, Y.; Huang, M.; Garcia, A.; Hernandez, A.; Song, H.; Comput. Electron. Agr. 58 (2007) 144.

[39] Dalal, R. Henry, R., Soil Sci. Soc. Am. J. 50 (1986) 120.

[40] Chang, C.; Laird, D.; Soil Science 167 (2002) 110.

[41] Madari, B.; Reeves, J.; Machado, P.; Torres, E.; McCarty, G.; Geoderma 136 (2006) 245.

[42] Barthés, B.; Brunet, D.; Ferrer, H.; Chotte, J.; Feller, C.; J. Near Infrared Spectrosc. 14 (2006) 341.

[43] http://www.dnpm.gov.br/assets/galeriadocumento/balancomineral2001/gipsita.pdf, Acessado em 08/11/2010

[44] Suykens., J.A.K.; Pelckmans, K.; Van Gestel, T., de Brabanter, J., Lukas, L., Hamers, B.; de Moor, B.; Vandewalle, J., LS-SVMlab Toolbox User's Guide version 1.5, Departament of Electrical Engineering, Katholieke Universiteit Leuven, 2003.

[45] Kennard, R.W.; Stone, L.A. ; Technometrics 11 (1969) 137.

[46] Sekulic, S.; Walkeman, J., Doherty, P., Hailey, P.A., J.Pharm. Biomed. Anal. 17 (1998) 1285.

[47] Berntsson, O.; Danielsson, L-G; Folestad, S.; Anal. Chem. Acta 364 (1998) 243.

[48] Berntsson, O.; Danielsson, L-G; Johansson; M.O., Folestad, S.; Anal. Chem. Acta 419 (2000) 45.

[49] Berntsson, O.; Danielsson, L-G; Lagerholm, B.; Folestad, S.; Powder Tecnol. 123 (2002) 185.

[50] Burns, D.; Ciurczak, E.; Handbook of near-infrared analysis, Marcel Dekker, New York, 1992.

[51] Rantanen, J.; Lehtola, S.; Ramet, P.; Mannermaa, J.; Yliruusu, J.; Powder Tecnol. 99 (1998) 1998.

[52] Soon, M.; Pat, G.F.; J.Pharm. Biomed. Anal. 14 (1996) 1681.

[53] Jedvert, I.; Josefson, M.; Langkilde, F.; J. Near Infrared Spectrosc. 6 (1998) 279.

[54] Rawlings, D.; Microb Cell Fact 4 (2005) 1.

[55] Jerez, C.; FEMS Microbiol Lett 56 (1988) 289.

[56] Xiao, S.; Chao, J.; Wang, W.; Fang, F.; Qiu, G.; Liu, X.; Folia Biol 55 (2009) 1.

[57] Hubert, W.A.; Leduc, L.; Ferroni, G.D.; Curr Microbio 31 (1995) 10.

[58] Modak, J.M.; Natarajan, K.; Mukhopadhyay, S; Hydrometallurgy 42 (1996) 51.

[59] Seeger, M; Jerez, C.A.; FEMS Microbiol Rev 11 (1993) 37.

[60] Seeger, M.; Osorio, G.; Jerez C.A.; FEMS Microbiol Lett 138 (1996) 129.

[61] Farah, C.; Vera, M.; Morin, D.; Haras, D.; Jerez, C.A.; Guiliani, N; Appl Environ Microbiol 71 (2005) 7033.

[62] Norris, K.P.; Hygiene 57 (1959) 326.

[63] Naumann, D.; Fijala, V.; Labischinski, H.; Giesbrecht, P.; Modern techniques for rapid microbiological analysis., VHC publishers Inc., New York, 1991.

[64] Yu, C.; Irudayaraj, J.; Biopolymers 77 (2004) 368.

[65] Garcia Junior, O.; Rev Bras Microbiol 22 (1991) 1.

[66] Gidman, E.; Goodacre, R.; Emmet, B., Smith, A.R.; Gwynn-Jones, D.; Phytochemistry 63 (2003) 705.

[67] Sharma, P.; Das A, H.; Hydrometallurgy 71 (2003) 285.

[68] Vígh, L., Maresca, B.; Harwood, J.L.; Trends Biochem Sci 23 (1998) 369.

[69] Brandenburg, K.; Seydel, U.; Eur. J. Biochem. 191 (1990) 229.

[70] Zhou, H.; Gu, G.W.; Chin. J. Digest. 8 (1998) 10.

[71] Anthony, P.P.; Histopathology 39 (2001) 109.

[72] Kojiro, M.; Hepatol Res 37 (2007) 121.

[73] International Working Party, Hepatology 22 (1995) 983.

[74] Wayne, J.D.; Lauwers, G. Y.; Ikai, I.; Doherty, D. A.; Belghiti, J., Yamaoka, Y.; Regimbeau, J.; Nagorney, D. M.; Do, K.; Ellis, L. M.; Curley, S. A.; Pollock, R. E.; Vauthey, J.; Ann Surg 235 (2002) 722.

[75] Scheuer, P.; Lefkowitch, J.H.; Liver biopsy interpretation, London, 2000.

[76] Kim, S.B.; Temiyasathit, C.; Bensalah, K.; Tuncel, A.; Cadeddu, J.; Kabbani, W.; Mathker, A.V.; Liu, H.; Expert Syst Appl 37 (2010) 3863.

[77] Honara, A.L.; Kangb, K.A.; Comp. Biochem. Phys. A 132 (2002) 9.

[78] Conti, C.; Ferraris, P.; Giorgini, E.; Rubini, C.; Sabbatini, S.; Tosi, G.; Anastassopoulou, J.; Arapantoni, P.; Boukaki, E.; Konstadoudakis, S.; Theophanides, T.; Valavanis, C.; J Mol Struct 881 (2008) 46.

[79] Clarke, F.; Hammond, S.V.; Jee, R.D.; Moffat, C.A.; Appl. Spectrosc. 56 (2002) 1475.

[80] Chalmers, J. M., Griffiths, P.R.; Handbook of Vibrational Spectroscopy, vol 2, John Wiley & Sons, New York, 2002.