UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE ALIMENTOS
CURSO DE ENGENHARIA DE ALIMENTOS

GUSTAVO YASUO FIGUEIREDO MAKIMORI

DESENVOLVIMENTO DE MÁQUINAS DE VETOR SUPORTE PARA A CLASSIFICAÇÃO DE CAFÉ ARÁBICA VERDE POR ESPECTROSCOPIA DE INFRAVERMELHO MÉDIO

TRABALHO DE CONCLUSÃO DE CURSO

CAMPO MOURÃO
2015

GUSTAVO YASUO FIGUEIREDO MAKIMORI

DESENVOLVIMENTO DE MÁQUINAS DE VETOR SUPORTE PARA A CLASSIFICAÇÃO DE CAFÉ ARÁBICA VERDE POR ESPECTROSCOPIA DE INFRAVERMELHO MÉDIO

Trabalho de conclusão de curso de graduação, apresentado ao Curso Superior de Engenharia de Alimentos do Departamento Acadêmico de Alimentos, da Universidade Tecnológica Federal do Paraná – UTFPR, Câmpus Campo Mourão, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Alimentos.

CAMPO MOURÃO

2015

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Departamento Acadêmico de Alimentos

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

DESENVOLVIMENTO DE MÁQUINAS DE VETOR SUPORTE PARA A CLASSIFICAÇÃO DE CAFÉ ARÁBICA VERDE POR ESPECTROSCOPIA DE INFRAVERMELHO MÉDIO

POR

GUSTAVO YASUO FIGUEIREDO MAKIMORI

Trabalho de Conclusão de Curso (TCC) apresentado em 02 de julho de 2015 às 14:00 horas como requisito parcial para obtenção do título de Bacharel em Engenharia de Alimentos. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho APROVADO.

_____
Prof°. Dr°. Evandro Bona
Orientador

_____
Profª. Drª. Ailey Aparecida Coelho Tanamati
Membro da banca

_____
Profº. Drº. Paulo Henrique Março
Membro da banca

Nota: O documento original e assinado pela Banca Examinadora encontra-se na Coordenação do Curso de Engenharia de Alimentos da UTFPR Câmpus Campo Mourão.

AGRADECIMENTOS

Aos meus pais, Edson Yasuo Makimori e Rosalira Figueiredo Makimori, pelos valores ensinados, por toda compreensão, apoio e incentivo.

Ao Professor Dr. Evandro Bona pela dedicada orientação, pelos conselhos e confiança para a realização deste trabalho.

Ao Professor Dr. Heron Oliveira Santos Lima pela franqueza e sinceridade em suas orientações.

Ao Professor Dr. Charles Windson Isidoro Haminiuk pela iniciação científica realizada com sucesso.

Aos familiares de Campo Mourão e Campo Grande pelo suporte e incentivo.

Aos amigos André Luis Guimarães Lemes, Rodrigo Mochi Guazelli e Alexandre Guimarães Inácio pela amizade e parceria que levarei sempre comigo

Aos amigos de classe Fernanda Rubio, Tânia Barbedo, Alini Gomes, Rayssa Simoni, Marília Gato, Mariana Terao, Eduardo Esperança, Matheus Vicente, Ana Gabriela Anthero, Amanda Salgado, Mateus De Souza, Isadora Tavares, Paula Rosa, Nathália Mercante e Tamires da Silva pelo companheirismo nesta caminhada.

À família que se tornou o condomínio Santa Cecília durante todo o curso.

RESUMO

MAKIMORI, Gustavo Yasuo Figueiredo. **Desenvolvimento de máquinas de vetor suporte para a classificação de café arábica verde por espectroscopia de infravermelho médio.** 2015. 33 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2015.

O Brasil é o maior produtor e exportador de café do mundo sendo uma importante *commodity* econômica do país. As duas espécies de café com maior valor econômico são o *canephora e o arábica,* sendo o último considerado de maior valor econômico por gerar uma bebida de melhor qualidade. Clima, espécie, método de cultivo e industrialização também são determinantes para a qualidade final da bebida. O objetivo deste trabalho foi desenvolver uma metodologia que seja capaz de discriminar genótipos de café arábica verde e também sua origem de plantio utilizando espectroscopia de infravermelho médio com transformada de Fourier (FTIR) e máquinas de vetor suporte (SVM, do inglês *support vector machine*). Para tanto foram coletados espectros FTIR de 74 amostras de 20 genótipos diferentes plantados nas cidades de Paranavaí, Cornélio Procópio, Mandaguari e Londrina. Para analisar os espectros foram construídas SVMs usando bases radiais como funções *kernel* e a estratégia *one-against-all* como abordagem multiclasses. As SVMs desenvolvidas tiveram sua eficiência avaliada através da sensibilidade e especificidade para as amostras de teste. Quanto à origem geográfica as amostras foram satisfatoriamente classificadas com uma sensibilidade média de 97,5% e especificidade média de 99,4%. Já para a classificação genotípica o desempenho não foi adequado com uma sensibilidade média de 66,0% e uma especificidade de 95,6%. Além disso, a classificação geográfica demonstrou-se mais fácil, pois menos amostras foram selecionadas como vetores suporte. O desequilíbrio na quantidade de amostras para o problema de classificação por genótipo pode ser a causa da baixa sensibilidade da SVM. Assim, sugere-se a busca de outras abordagens de problemas multiclasse para o aperfeiçoamento dos modelos propostos.

Palavras-chave: FTIR. Reconhecimento de padrões. SVM.

ABSTRACT

MAKIMORI, Gustavo Yasuo Figueiredo. **Development of support vector machines for green arabica coffee classification by mid-infrared spectroscopy.** 2015. 33 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2015.

Brazil is the world's largest producer and exporter of coffee being an important economic commodity in the country. The two species of greatest economic value are *canephora* and *arabica*, being the last one considered of greater economic value by generating a better quality beverage. Climate, species, cultivation method and industrialization are also critical for the final quality of the beverage. The objective of this study was to develop a methodology that is capable to discriminate different green *arabica* coffee genotypes and also their geographical origin by using mid-infrared spectroscopy with Fourier transform (FTIR) and support vector machines (SVM). Therefore, 74 FTIR spectra were collected from 20 different genotypes planted in the cities of Paranavaí, Cornélio Procópio, Mandaguari and Londrina. To analyze the spectra were built SVMs using radial basis as kernel function and the one-against-all multiclass approach. The developed SVM were evaluated by sensitivity and specificity for the test samples. For the geographic origin the samples were successfully classified with an average sensitivity of 97.5% and average specificity of 96.9%. Otherwise, for genotypic classification the performance was not satisfactory with an average sensitivity of 66.0% and a specificity of 95.6%. Furthermore, the geographical classification proved to be easier because fewer samples were selected as support vectors. The unbalance in the number of samples for genotype classification problem can be the cause of poor sensitivity of the SVM. Thus, it is suggested to search for other approaches to multiclass problems for the improvement of the proposed models.

Keywords: FTIR. Pattern Recognition. SVM.

Sumário

1. INTRODUÇÃO

O Brasil destaca-se como o maior produtor e exportador de café do mundo. Com um volume recorde de exportação no atual ano safra, apesar da recente retração econômica, o valor supera em 12,4% quando comparado ano anterior (ICO, 2015). A produção brasileira, de aproximadamente 45 milhões de sacas, superou a do Vietnã e Colômbia (respectivos maiores produtores) somados, no ano de 2014 (ABIC, 2015). A indústria cafeeira é responsável de forma direta e indireta por mais de oito milhões de empregos, assim, compreende-se a importância de tal *commoditie* na economia do país (MAPA, 2015).

O café possui características botânicas de árvores e arbustos tropicais relacionadas com o gênero *Coffea* da família *Rubiaceae*. As espécies mais produzidas devido ao valor comercial são: *arabica* (arábica) e *canephora* (robusta ou conilon). Por gerar uma bebida de notas sensoriais mais agradáveis, o café arábica possui valor comercial mais alto que o robusta (CESARINO & MAZZAFERA, 2015).

O aroma e sabor complexo da bebida gerada pela torra do grão de café é proveniente de diversos fatores que incluem o genótipo, clima, localização geográfica, métodos de cultivo, armazenamento e industrialização do grão (KLEINWÄCHTER *et al.*, 2015). Mais recentemente, estudos demonstram que o consumo moderado da bebida está ligado a redução de fibrose hepática decorrente de hepatite C (MACHADO; PARISE; de CARVALHO, 2014). Relata-se também que compostos antioxidantes presentes no café possuem propriedades neuroprotetoras, podendo auxiliar no combate de doenças neurodegenerativas como Alzheimer, Parkinson e Isquemia, pois previnem a morte de células neurais (KIM *et al*., 2005).

Quando ainda verdes, os grãos de espécies diferentes são de fácil identificação por um técnico treinado. Porém, após o processo de moagem e torrefação, identificar e assegurar que o espécime em questão não foi adulterado exige técnicas mais específicas (KEMSLEY; RUAULT; WILSON,1995). Além disso, a identificação por genótipo e local de cultivo é impossível de ser realizada por uma simples inspeção visual (LINK, 2014).

A espectroscopia de infravermelho médio com transformada de Fourier (FTIR) tem se demonstrado uma técnica simples e rápida que proporciona uma impressão (*fingerprint*) do espécime em questão (WANG e LIM, 2012). A radiação é emitida na região de número de onda de 4000 a 400 cm^{-1} e separada em dois feixes, sendo um fixo e outro móvel. Com a variação das distâncias percorridas pelos dois feixes, obtêm-se uma sequência de interferências que geram variações na intensidade de radiação recebida pelo detector, chamado de interferograma. A transformação de Fourier em posições sucessivas do espelho dá origem ao espectro completo de infravermelho. Como a técnica permite uma alta resolução do espectro e utiliza uma grande faixa de comprimento de onda, pode se obter uma quantidade enorme de variáveis (SILVERSTEIN; WEBSTER; KIEMLE, 2007). Devido a natureza multivariada desses espectros, métodos quimiométricos são indispensáveis para o tratamento correto destas informações (PARREIRA, 2003).

As redes neurais artificiais (RNA) são um conjunto de métodos matemáticos multivariados que podem ser aplicados através de algoritmos computacionais. Seu funcionamento é baseado no cérebro humano e ganhou espaço nos últimos anos em aplicações de reconhecimento de padrões devido sua capacidade de armazenamento de novos dados e generalização (HAYKIN, 2001). Mais recentemente, as máquinas de vetor suporte (SVM, do inglês *suport vector machine*) têm se demonstrado uma ferramenta interessante (FERRAO *et al*., 2007). Seu algoritmo pode ser utilizado para classificação de padrões e regressão. O funcionamento de uma SVM tem como ideia principal a construção de um hiperplano como superfície de separação onde a margem de decisão seja máxima entre exemplos positivos e negativos (HAYKIN, 2001). Durante a construção de uma SVM o algoritmo indutivamente controla a complexidade do modelo independentemente da dimensionalidade do problema em questão. A SVM leva vantagem quando comparada com outros modelos matemáticos de aproximação como, por exemplo, as redes neurais, o que evita sobre ajuste do modelo e maior capacidade de generalização (MARETTO, 2011).

Recentemente, relatou-se o uso de SVM na discriminação geográfica de azeites italianos (DEVOS *et al*., 2014), sólidos não gordurosos em leite cru (BASSBASI *et al*., 2014) e classificação de parâmetros de qualidade em suco de morango (vitamina C, pH, sólidos solúveis totais, acidez total e taxa de açúcar/ácido) (QIU, *et al*., 2014). Neste contexto, teve-se por objetivo testar a aplicação das SVMs na classificação geográfica e genotípica de cafés brasileiros.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Desenvolver uma metodologia que seja capaz de discriminar diferentes genótipos de café arábica verde e também sua origem de plantio utilizando FTIR e SVM.

2.2. OBJETIVOS ESPECÍFICOS

- Coletar, registrar e armazenar as amostras dos genótipos que serão fornecidas pelo Instituto Agronômico do Paraná (IAPAR, Londrina – PR);
- Obter os espectros infravermelhos no equipamento de FTIR e realizar os pré-processamentos necessários (correção de linha de base, normalização, suavização, análise de componentes independentes - ICA, etc.);
- Testar as diferentes formas de apresentação dos espectros (espectro puro, primeira derivada, segunda derivada);
- Encontrar os melhores parâmetros para as SVM e avaliar sua capacidade de classificação correta.

3.METODOLOGIA

3.1. AMOSTRAS

O Instituto Agronômico do Paraná com sede em Londrina forneceu 74 amostras de 20 genótipos de café arábica plantados nas cidades de Mandaguari (MD), Cornélio Procópio (CP), Londrina (LD) e Paranavaí (PV) das safras de 2009 e 2010 (Tabela 1). Todas as amostras são de grãos verdes secos, moídos e devidamente embalados.

Tabela 1. Relação de amostras fornecidas pelo IAPAR - Londrina.

Genótipo	Ano	Local	Amostras	Genótipo	Ano	Local	Amostras
IP097	2009	Mandaguari	1	IP105	2009	Mandaguari	1
	2010	Paranavaí	2		2010	Paranavaí	1
		Mandaguari	1			Cornélio Procópio	1
		Londrina	1			Mandaguari	1
IP098	2009	Mandaguari	1			Londrina	1
	2010	Mandaguari	1	IP106	2009	Mandaguari	1
		Londrina	1		2010	Paranavaí	1
IP099	2009	Mandaguari	1			Cornélio Procópio	1
	2010	Paranavaí	1			Mandaguari	1
		Cornélio Procópio	1			Londrina	1
		Mandaguari	1	IP107	2009	Mandaguari	1
		Londrina	1		2010	Mandaguari	1
IP100	2009	Mandaguari	1			Londrina	1
	2010	Paranavaí	1	IP108	2009	Mandaguari	1
		Mandaguari	1		2010	Paranavaí	1
		Londrina	1			Mandaguari	1
IP101	2009	Mandaguari	1			Londrina	1
	2010	Cornélio Procópio	2	CT001	2009	Mandaguari	1
		Mandaguari	1		2010	Paranavaí	2
		Londrina	1			Mandaguari	1
IP102	2009	Mandaguari	1			Londrina	1
	2010	Paranavaí	1	BB001	2009	Mandaguari	1
		Mandaguari	1		2010	Cornélio Procópio	1
		Londrina	1	TU001	2010	Paranavaí	3
IP103	2009	Mandaguari	1	IA059	2009	Mandaguari	1
	2010	Mandaguari	1		2010	Paranavaí	2
		Londrina	1			Cornélio Procópio	2
IP104	2009	Mandaguari	1			Mandaguari	1
	2010	Paranavaí	1			Londrina	1
		Mandaguari	1	IC001	2009	Mandaguari	1
		Londrina	1	MN001	2010	Cornélio Procópio	2
IE105	2010	Paranavaí	2	IE059	2010	Paranavaí	2

3.2. ESPECTROSCOPIA DE INFRAVERMELHO MÉDIO COM TRANSFORMADA DE FOURIER (FTIR)

Com o auxílio de uma prensa hidráulica (Bovenau, P15 ST) pastilhas translúcidas foram formadas, em quintuplicatas, contendo 100mg de KBr (padrão cromatográfico) e 1 mg de amostra em molde padrão (ICL, ICL's Macro/Micro KBr die) sob 7 toneladas de pressão. Primeiramente foi realizada uma leitura sem amostra (*background*) com o intuito de descontar a presença do ar nos demais espectros. O FTIR (Shimadzu, IR Affinity-1) monitorou a região de 4000 a 400 cm^{-1} para cada pastilha sendo aplicada uma apodização do tipo Happ-Genzel com 32 varreduras acumuladas para a formação do espectro final. Para a SVM foi utilizada apenas a região entre 1900 e 800 cm^{-1} onde se encontram os compostos químicos mais importantes para a caracterização de amostras de café (KEMSLEY; RUAULT; WILSON,1995; WANG e LIM, 2012; LINK *et al*., 2014).

3.3. PRÉ-TRATAMENTO

Antes da utilização dos espectros nos modelos de classificação alguns tratamentos foram realizados. Utilizando o software IRsolution 1.5 (Shimadzu Corporation, Kyoto, Japão) presente no equipamento, foi feita a correção de linha de base e a suavização. Já a normalização, primeira e segunda derivada foram implementadas no MATLAB R2008b através do algoritmo de Savitzky-Golay (SAVITZKY; GOLAY, 1964; WANG *et al*., 2009)

O efeito de cada pré-tratamento (espectros puros, 1ª derivada e 2ª derivada) foi avaliado através da eficiência de classificação obtida pela SVM.

3.4 MÁQUINA DE VETOR SUPORTE

Em 1995, *Vapnik-Chervonenkis* propuseram as SVM baseadas na teoria de minimização estrutural de risco (SRM, do inglês *structural risk minimization*). A SRM tem como princípio que a taxa de erro nas amostras de teste (erro de generalização) é limitada pela soma da taxa de erro de treinamento e por um termo dependente da dimensão de *Vapnik-Chervonenkis* (V-C). Sendo, a dimensão V-C o número máximo de exemplos de treinamento que podem ser aprendidos pela máquina sem erro. Portanto, para dados linearmente separáveis e independentes entre si, modelos desenvolvidos com este princípio tem maior capacidade de generalização para amostras desconhecidas (HAYKIN, 2001).

Para o problema multiclasses proposto existem dois tipos principais de construção da SVM, *one-against-one* (OAO) e *one-against-all* (OAA). O modelo desenvolvido foi OAA, ou seja, o modelo separa uma classe de todas as demais, diferentemente da OAO que constrói modelos separando uma classe contra outra (LI *et al*., 2009). O modelo OAA foi escolhido devido à maior facilidade e rapidez de aplicação e resultados tão apurados quanto outras metodologias multiclasse (RIFKIN e KLAUTAU, 2004). Logo, no caso da classificação geográfica, 4 SVMs foram desenvolvidas (4 cidades). Com um total de 364 amostras, 80% deste valor foi utilizado para o treinamento das SVM e 20% para teste. Portanto, na classificação geográfica foram utilizadas 291 amostras para treinamento e 73 amostras para teste.

Já na classificação genotípica, na tentativa de balancear o banco de dados e promover uma melhora na classificação ocorreu uma seleção prévia dos genótipos. O critério de seleção foi utilizar os genótipos que possuíam ao menos 4 amostras conforme a Tabela 1. Assim, na classificação genotípica foram construídas 11 SVMs utilizando 208 amostras para treinamento e 52 amostras para teste.. A construção de uma SVM está baseada em dois princípios, a construção de um hiperplano de separação ótimo e a utilização de funções *kernel* para tratar dados que não são linearmente separáveis (LI *et al*., 2009).

A ideia da construção de um hiperplano onde a margem decisória seja máxima para um exemplo binário pode ser observada na Figura 1.

Observa-se que há várias fronteiras decisórias, porém a reta vermelha, encontrada utilizando-se a teoria SRM, separa as duas classes com maior distância possível entre as diferentes amostras. Nota-se que para o caso, apenas dois vetores suportes (SVs, do inglês *support vector*) de cada classe foram necessários para a construção das margens decisórias (LIMA, 2004).

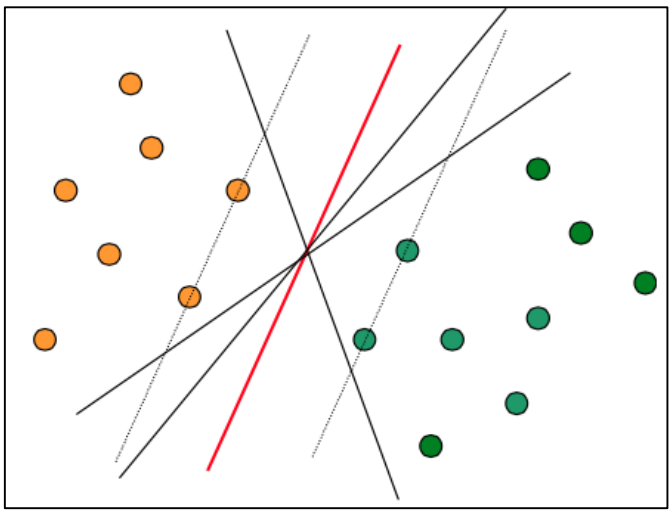


Figura1. Construção do hiperplano ótimo.
Fonte. Lima (2004).

A Equação 1 rege a superfície decisória na forma do hiperplano

$$w^T x + b = 0 \qquad (1)$$

onde *x* é o vetor de entrada, *b* é o bias, *w* é um vetor peso ajustável que fornecerá a máxima separação entre os vetores suporte minimizando a norma euclidiana do mesmo. Encontrar os melhores valores de *w* pressupõem a solução do problema primordial. Este consiste nas restrições lineares de *w*, porém do ponto de vista computacional, a função de erro é quadrática e portanto convexa. A solução desse problema de otimização quadrática com restrições é feita aplicando-se o método dos multiplicadores de Lagrange (HAYKIN, 2002).

A não linearidade dos dados é algo comum, logo uma saída interessante para tornar os dados lineares é a utilização de uma função *kernel*. O método é baseado no aumento das dimensões, logo há um

distanciamento das amostras e hiperplanos podem ser gerados separando as classes (Figura 2) (LI *et al*., 2009).

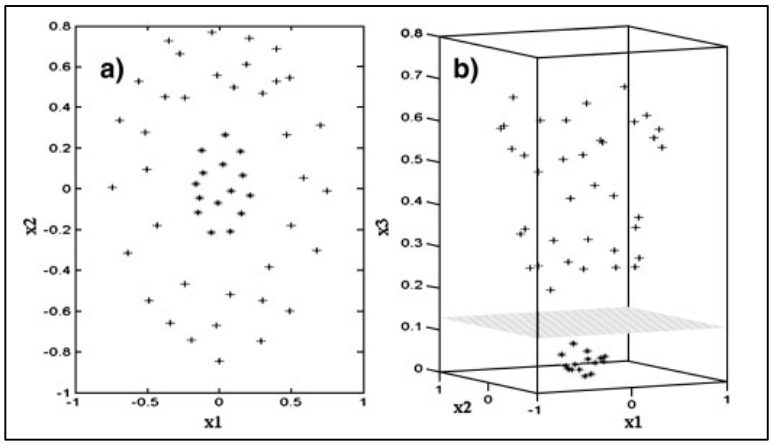


Figura 2. Gráfico (a) com duas dimensões e sua respectiva transformação para um espaço de características com 3 dimensões (b) onde foi possível a separação por um plano. Fonte. Li et al., (2009).

A exemplo do caso, supondo que uma amostra possua coordenada $x_i = [x_{i1}, x_{i2}]$. O cálculo da terceira coordenada pode ser calculado simplesmente por $x_{i3} = {x_{i1}}^2 + {x_{i2}}^2$. Portanto, para um espaço de 3 dimensões teremos $x_i = [x_{i1}, x_{i2}, x_{i3}]$ (LI *et al*., 2009).

As funções do tipo *kernel* projetam os dados em um espaço superior de dimensões baseado no cálculo do produto interno ($K = K_{ij} = K(x_i, x_j)$). Para ser uma função *kernel* a mesma deverá respeitar o Teorema de Mercer, ou seja, produzir matrizes com autovalores maiores que zero. Dentre as funções que obedecem o Teorema de Mercer destacam-se as máquinas de apredizagem polinomial, perceptron de duas camadas e rede de função de base radial (HAYKIN, 2002). A função *kernel* utilizada foi do tipo função de base radial (Equação 2), pois trabalha com um único parâmetro (γ), que está relacionado com a suavidade da função, facilitando a seleção de um valor adequado para o mesmo para a aplicação desejada.

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (2)$$

onde (x_i e x_j) são dois vetores diferentes e γ é um parâmetro de controle escolhido *a priori*.

Para dados que não são linearmente separáveis é acrescentado um parâmetro C, escolhido *a priori*, ao problema de otimização de w. O parâmetro C controla o compromisso entre a complexidade da máquina e o número de padrões não separáveis (HAYKIN, 2001). Assim, a função sinal (sng) decisória da SVM pode ser descrita pela Equação 3.

$$sng(w^T x + b) = sng\left[\sum_{i=1}^{l} y_i\, \alpha_i\; K(x_i, x_j) + b\right] \tag{3}$$

onde *x* é um vetor de entrada, *w* é um vetor peso ajustado, y é um vetor indicador (vetor de saída) que $\in \{1, -1\}$, $\propto$ são os multiplicadores de Langrange e está contido entre $0 \leq \alpha \leq \mathrm{C}$, $K(x_i, x_j)$ é a função kernel descrita pela Equação 2 e *b* é o *bias.* Portanto, a arquitetura da SVM por ser esquematizada conforme Figura 3. Uma abordagem mais abrangente sobre as SVM pode ser encontrada em Chang & Lin (2011).

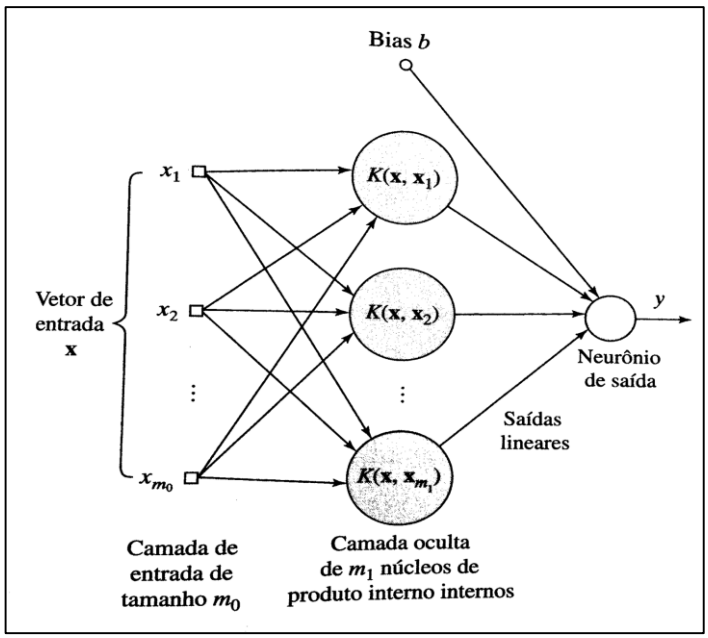


Figura 3. Arquitetura da SVM.
Fonte. Haykin (2002).

3.5 IMPLEMENTAÇÃO COMPUTACIONAL

Todos os algoritmos dos procedimentos matemáticos supracitados foram realizados pelo software MATLAB R2008b. As rotinas utilizadas para as SVM pertencem ao repositório aberto LIBSVM e foram desenvolvidas por Chang & Lin (2011).

A eficiência da SVM depende de uma escolha correta dos parâmetros C e γ. Um *grid search* de varredura do log γ x log C foi realizado na tentativa de encontrar um par destes valores onde a porcentagem de classificação correta fosse mais alta. Uma primeira busca ocorreu entre -5 e 5 com variação de uma unidade. Após a definição de uma região subótima resultante, uma segunda varredura foi realizada usando uma variação de 0,5 unidades. O uso da escala logarítmica torna-se interessante pois cobre uma região maior de busca quando comparada com uma escala linear.

Para cada par de C e γ do *grid search*, as amostras de treinamento foram subdivididas em 10 grupos de validação cruzada. Nessa metodologia, as amostras de N-1 subgrupos são utilizadas no treinamento e o subgrupo restante utilizado para a validação. Tal procedimento é repetido N vezes onde ao final é calculada uma porcentagem média de classificação correta (BISHOP, 2002).

Após a definição dos parâmetros ótimos a SVM foi novamente construída para avaliar a capacidade de generalização nas amostras de teste. Nessa etapa as SVMs foram avaliadas através da sensibilidade e especificidade. A sensibilidade é capacidade do modelo de classificar amostras da classe como sendo da classe, já a especificidade reflete a capacidade de classificar as amostras que não pertencem a classe como não sendo da classe. O fluxograma da Figura 4 ilustra a implantação da metodologia descrita.

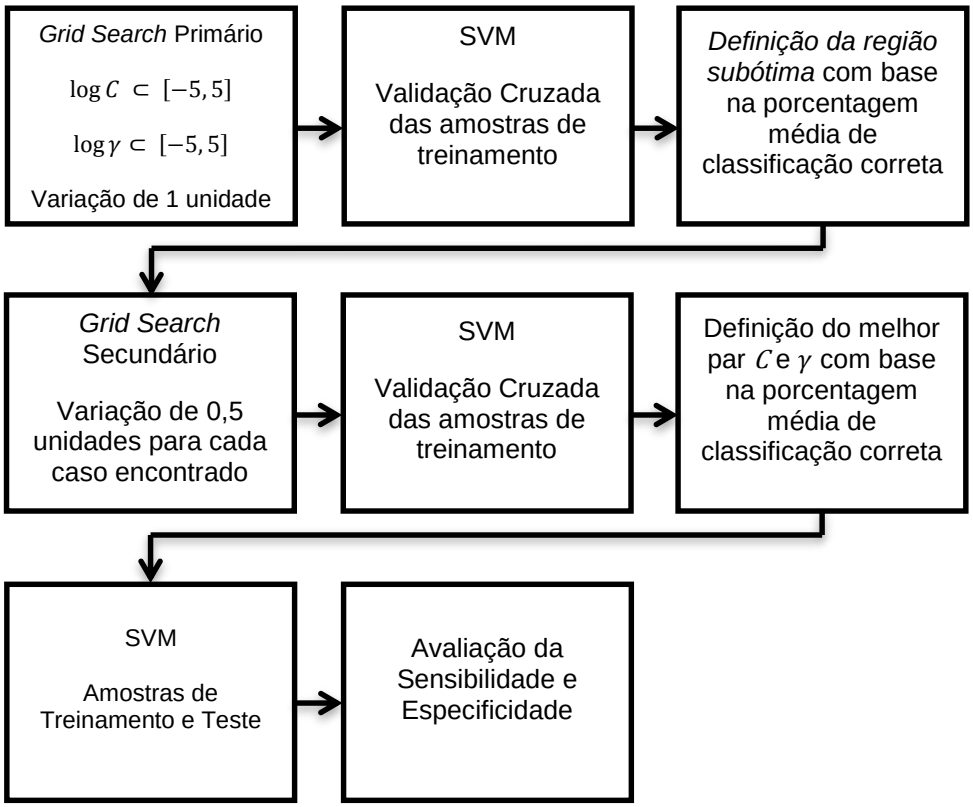
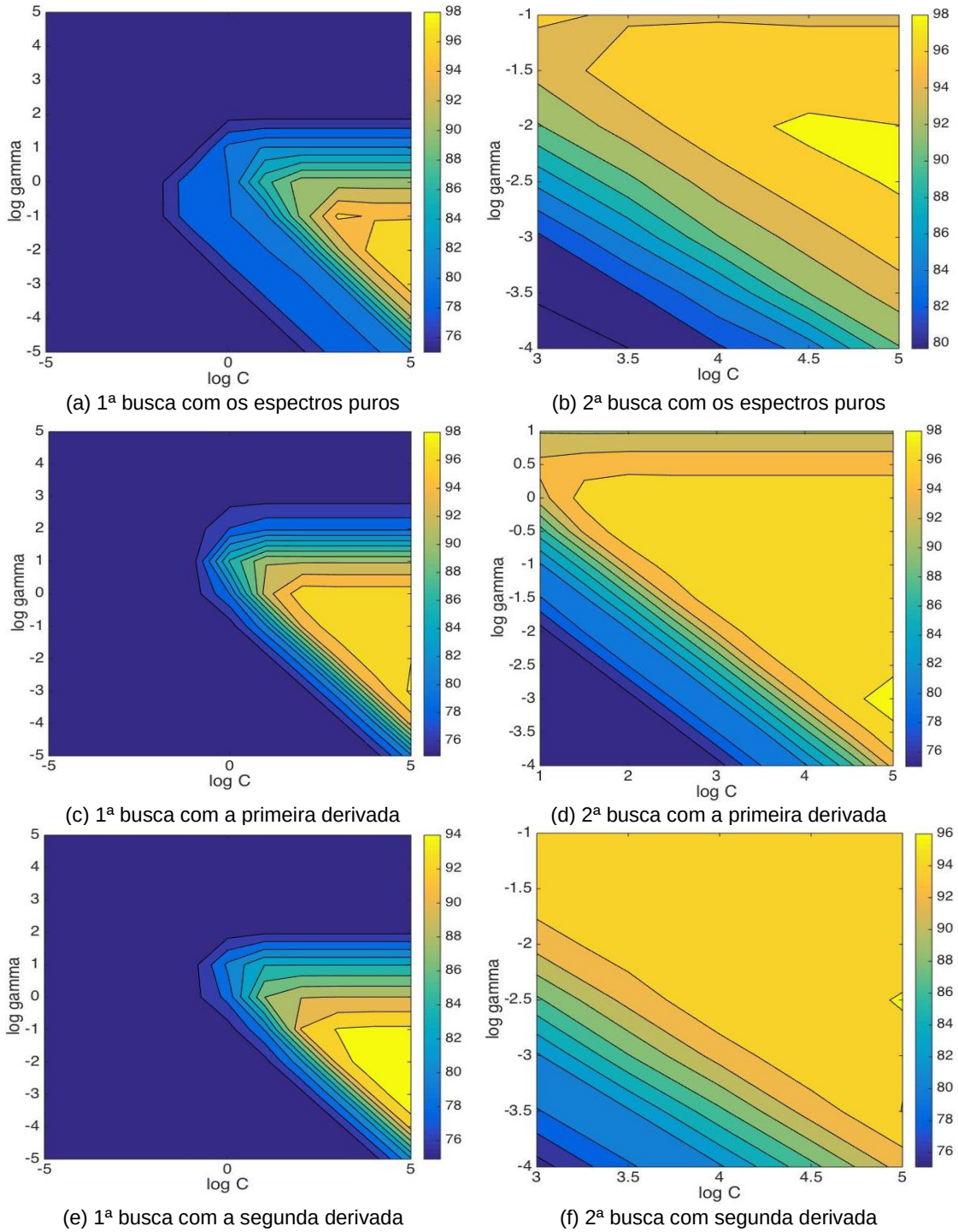


Figura 4. Fluxograma da metodologia aplicada.

4. RESULTADOS E DISCUSSÕES

4.1 CLASSIFICAÇÃO GEOGRÁFICA

A Figura 5 contêm os gráficos da primeira e segunda busca das melhores combinações entre log C x log γ.



(a) 1ª busca com os espectros puros

(b) 2ª busca com os espectros puros

(c) 1ª busca com a primeira derivada

(d) 2ª busca com a primeira derivada

(e) 1ª busca com a segunda derivada

(f) 2ª busca com segunda derivada

Figura 5. Os gráficos de (a) até (f) ilustram o processo de escolha do melhor par de C e γ para cada pré-tratamento testado. A escala de cores representa a porcentagem média de classificação correta.

Os melhores valores dos parâmetros C e γ estão dispostos na Tabela 2.

Tabela 2. Melhores parâmetros para otimização das SVM por cidade.

	Dados Puros	Primeira Derivada	Segunda Derivada
C	5,0118 x 10^{6}	1,0000 x 10^{5}	1,0000 x 10^{5}
$\boldsymbol{\gamma}$	2,5119 x 10^{-4}	1,0000 x 10^{-3}	3,1000 x 10^{-3}

Os valores de sensibilidade e especificidade da SVM usando os melhores parâmetros selecionados para cada pré-tratamento estão dispostos na Tabela 3.

Tabela 3. Sensibilidade e especificidade da SVM para a classificação por cidade.

Parâmetro	Cidade				Média
	Paranavaí	Cornélio Procópio	Mandaguari	Londrina	
Espectros Puros					
VS*	35	30	40	29	33,5
S (treinamento)**	1,0000	1,0000	1,0000	1,0000	1,0000
E* (treinamento)**	1,0000	1,0000	1,0000	1,0000	1,0000
S (teste)	1,0000	0,9000	1,0000	1,0000	0,9750
E (teste)	1,0000	1,0000	0,9773	1,0000	0,9943
Primeira Derivada					
VS	64	49	63	33	52,3
S (treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
E (treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
S (teste)	1,0000	0,9000	0,9655	0,9286	0,9485
E (teste)	1,0000	1,0000	0,9773	1,0000	0,9943
Segunda Derivada					
VS	73	52	71	47	60,8
S (treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
E (treinamento)	1,0000	1,0000	1,0000	1,0000	1,0000
S (teste)	1,0000	0,9000	0,9310	0,8571	0,9220
E (teste)	0,9811	1,0000	0,9773	1,0000	0,9896

* Quantidade vetores suporte selecionada para o modelo.
** Sensibilidade: capacidade do modelo de classificar amostras da classe como sendo da classe.
*** Especificidade: capacidade de classificar as amostras que não pertencem a classe como não sendo da classe.

A Figura 6 ilustra que a melhor SVM dentre os pré-tratamentos aplicados foi a alimentada com os espectros puros. A mesma conseguiu classificar as amostras de treinamento com 100% de sensibilidade e especificidade. Já para as amostras de teste a sensibilidade média foi de 97,5% e a especificidade média de 99,4%.

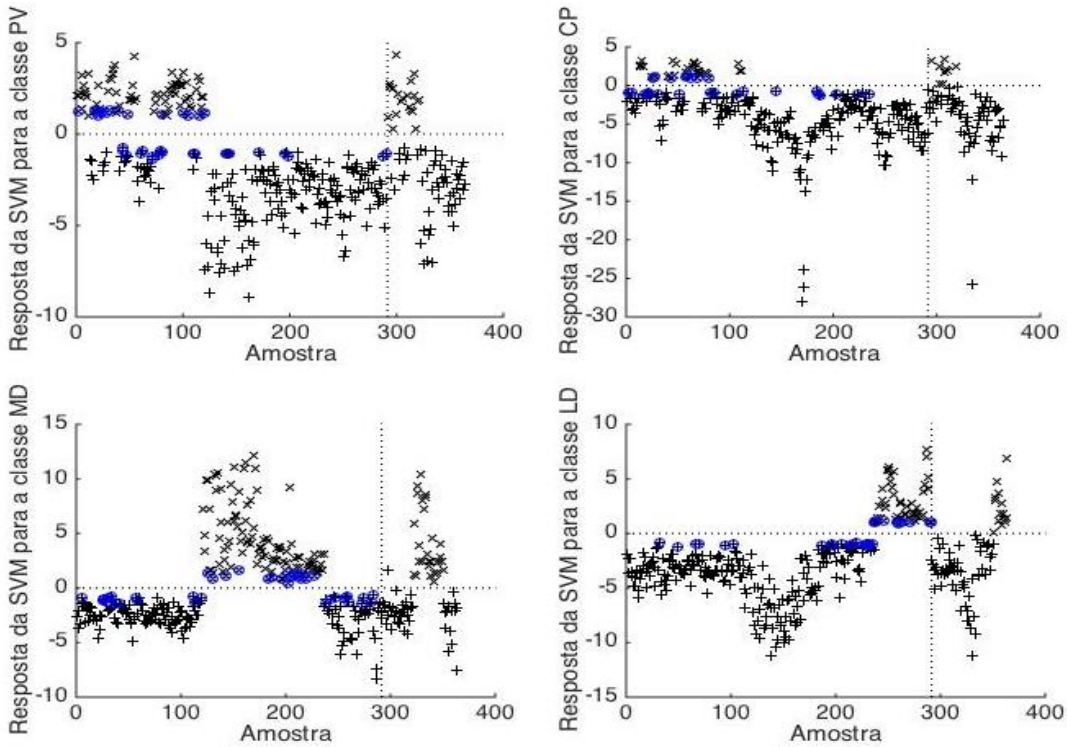
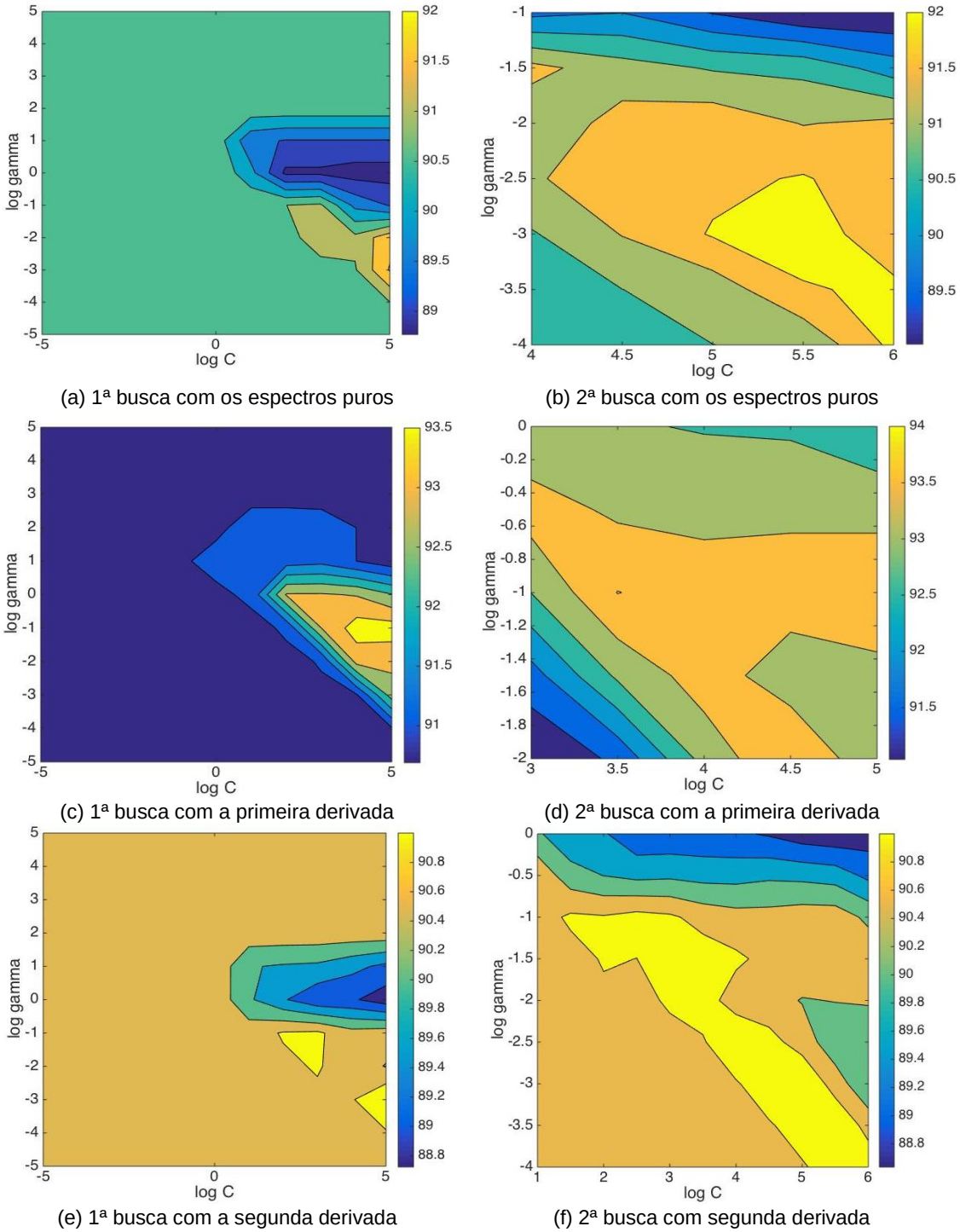


Figura 6. Resposta da melhor SVM. As amostras com sinal “x” são amostras da classe, sinal “+” não pertence a classe, circuladas em azul são as amostras selecionadas como SVs. A linha vertical tracejada separa as amostras de treinamento das de teste.

Em Lemes (2014) as mesmas amostras foram avaliadas usando um modelo PLS-DA (mínimos quadrados parciais com análise discriminante) construído com a primeira derivada e foi obtido, para as amostras de teste, uma sensibilidade média de 100% e uma especificidade média de 98,6%. O mesmo autor utilizou também uma rede neural artificial de base radial (RBF) alimentada com os *scores* do PLS-DA obtidos a partir dos espectros puros. Para esse último modelo foi obtida sensibilidade média, para as amostras de teste, de 99,1% e uma especificidade de 99,6%. As SVMs desenvolvidas obtiveram desempenho similar e contam com a vantagem de serem robustas e de fácil manuseio pois são utilizados os espectros sem a necessidade de uma redução prévia da dimensionalidade. O FTIR associado às SVMs demonstrou ser uma alternativa viável para a classificação geográfica de grãos verdes de café arábica.

4.2 CLASSIFICAÇÃO GENOTÍPICA

A Figura 9 contêm os gráficos da primeira e segunda busca das melhores combinações entre log C x log γ.



(a) 1ª busca com os espectros puros

(b) 2ª busca com os espectros puros

(c) 1ª busca com a primeira derivada

(d) 2ª busca com a primeira derivada

(e) 1ª busca com a segunda derivada

(f) 2ª busca com segunda derivada

Figura 9. Os gráficos de (a) até (f) ilustram o processo de escolha do melhor par de C e γ para cada pré-tratamento testado. A escala de cores representa a porcentagem média de classificação correta.

Os melhores valores de C e γ estão dispostos na Tabela 4.

Tabela 4. Melhores parâmetros para otimização das SVMs por genótipo

	Dados Puros	**Primeira Derivada**	**Segunda Derivada**
C	3,1622 x 10^{5}	3,1622 x 10^{3}	1,0000 x 10^{5}
$\boldsymbol{\gamma}$	1,0000 x 10^{-3}	1,0000 x 10^{-1}	1,0000 x 10^{-3}

Na Tabela 5, estão dispostos os dados relativos a sensibilidade e especificidade das SVM para a seleção de genótipos.

Tabela 5. Sensibilidade e especificidade da SVM para a classificação por genótipo.

Pré-Tratamento	Amostra	Parâmetro	Genótipo											Média
			IP100	IP102	IP104	IP105	IP106	IP108	IP097	IP099	CT001	IA059	IP101	
Espectro Puro	Treinamento	SV*	38	27	43	49	39	51	50	58	54	61	38	46,2
		S**	0,7143	1,0000	0,8000	0,5500	0,9500	0,6875	0,8000	0,5500	0,7000	0,7143	0,9500	0,7651
		E***	0,9948	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9889	1,0000	0,9985
	Teste	S	0,3333	1,0000	0,2500	0,2000	0,8000	0,5000	0,8000	0,0000	0,6000	0,1429	0,4000	0,4716
		E	0,9796	1,0000	1,0000	0,9574	0,9574	0,9792	0,8936	0,9362	0,9787	0,9556	1,0000	0,9655
1ª derivada	Treinamento	SV	40	33	42	51	43	55	46	56	48	53	37	45,8
		S	1,0000	1,0000	0,9333	0,9500	0,9000	0,9375	0,9500	0,9500	1,0000	0,9286	0,9500	0,9544
		E	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9947	1,0000	0,9947	1,0000	1,0000	0,9990
	Teste	S	0,3333	0,7500	0,5000	0,6000	0,6000	0,7500	0,8000	0,2000	0,6000	0,0000	0,8000	0,5394
		E	0,9796	0,9792	0,9583	0,9787	0,9787	1,0000	0,9362	0,9574	0,9574	0,9556	0,9787	0,9691
2ª derivada	Treinamento	SV	39	45	46	60	62	72	66	68	61	75	55	59,0
		S	0,7143	0,8667	0,8000	0,7000	0,8500	0,7500	0,8000	0,5000	0,9000	0,8571	0,9000	0,7853
		E	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9947	1,0000	1,0000	0,9995
	Teste	S	0,3333	0,2500	0,2500	0,2000	0,2000	0,7500	0,2000	0,2000	0,6000	0,1429	0,8000	0,3569
		E	0,9796	1,0000	0,9792	0,9787	0,9787	1,0000	1,0000	0,9787	0,9574	0,9111	0,9362	0,9727
1ª derivada com peso 10 para as amostras da classe	Treinamento	SV	44	33	51	48	39	52	46	53	50	68	38	47,5
		S	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
		E	1,0000	1,0000	1,0000	0,9947	1,0000	1,0000	0,9947	1,0000	0,9947	0,9889	1,0000	0,9975
	Teste	S	0,6667	0,7500	0,7500	0,8000	0,6000	0,7500	1,0000	0,4000	0,6000	0,1429	0,8000	0,6600
		E	0,9592	0,9792	0,9375	0,9574	0,9787	1,0000	0,8723	0,9574	0,9574	0,9333	0,9787	0,9556

* Quantidade vetores suporte selecionada para o modelo.
** Sensibilidade: capacidade do modelo de classificar amostras da classe como sendo da classe.
*** Especificidade: capacidade de classificar as amostras que não pertencem a classe como não sendo da classe.

Nenhum dos pré-tratamentos gerou resultados satisfatórios, porém para a primeira derivada, no geral, o desempenho foi melhor. Para as amostras de teste a SVM com a primeira derivada atingiu uma sensibilidade média de 53,94% e uma especificidade média de 96,91%. No problema de classificação genotípica cada SVM foi construída com, em média, 20 exemplos +1 (pertencentes à classe) e 200 exemplos -1 (não pertencentes à classe). Ou seja, trata-se de um problema de classificação desbalanceado resultando em um modelo de baixa sensibilidade. A Tabela 5 deixa claro que apesar de uma baixa sensibilidade, a especificidade atingiu valores satisfatórios. Para compensar o desequilíbrio na quantidade de exemplos de treinamento a SVM de melhor desempenho foi construída novamente usando-se um peso 10 vezes maior para os erros de classificação das amostras pertencentes à classe (CHANG & LIN, 2011). Assim, foi obtida uma sensibilidade média de 66,0% e uma especificidade de 95,6%.

As Figuras 10 e 11 ilustram os resultados obtidos pela melhor SVM para cada genótipo.

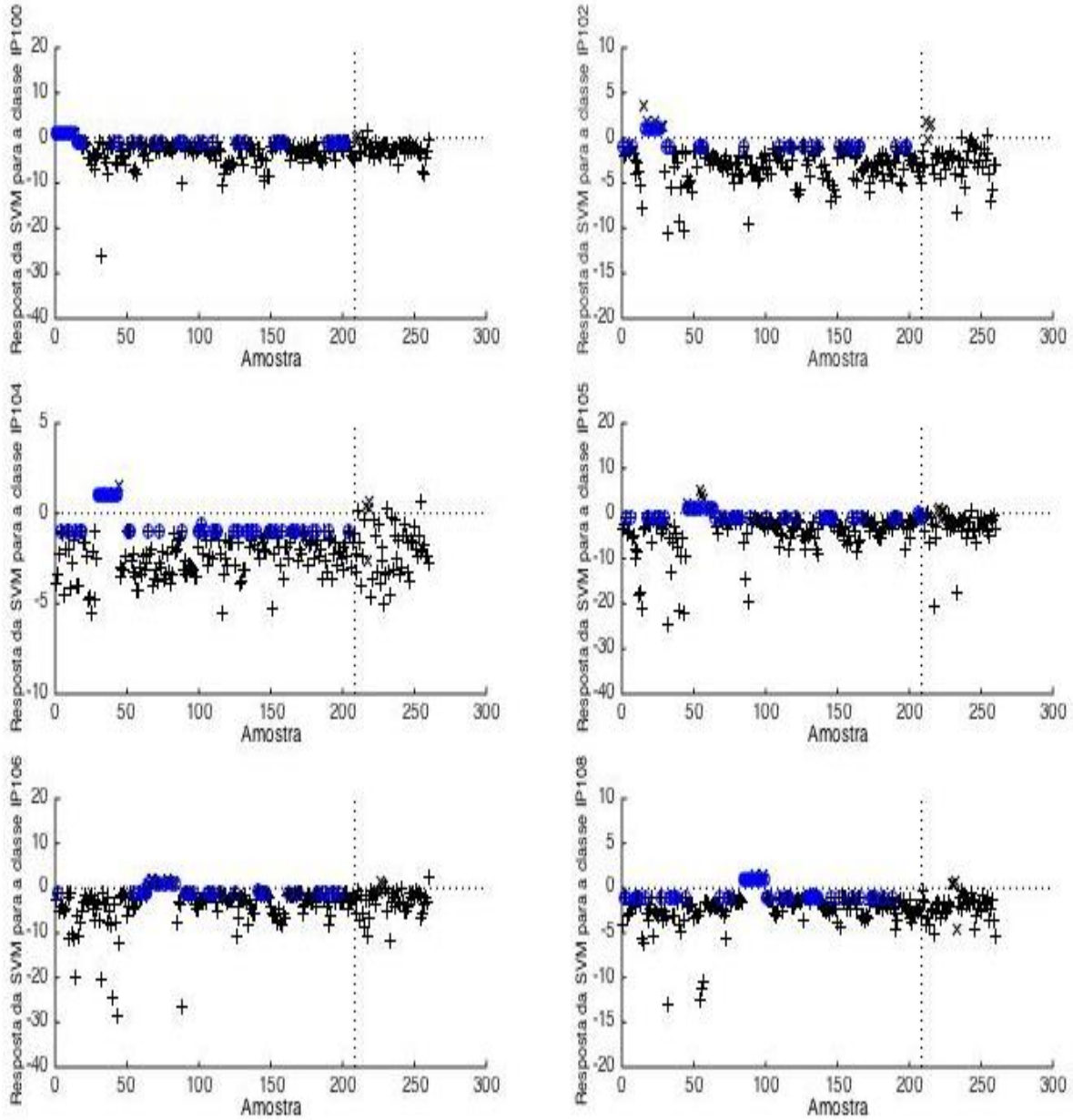


Figura 10. Resposta da melhor SVM. As amostras com sinal “x” são amostras da classe, sinal “+” não pertencem a classe, circuladas em azul são as amostras utilizadas como SVs. A linha vertical tracejada separa as amostras de treinamento das de teste.

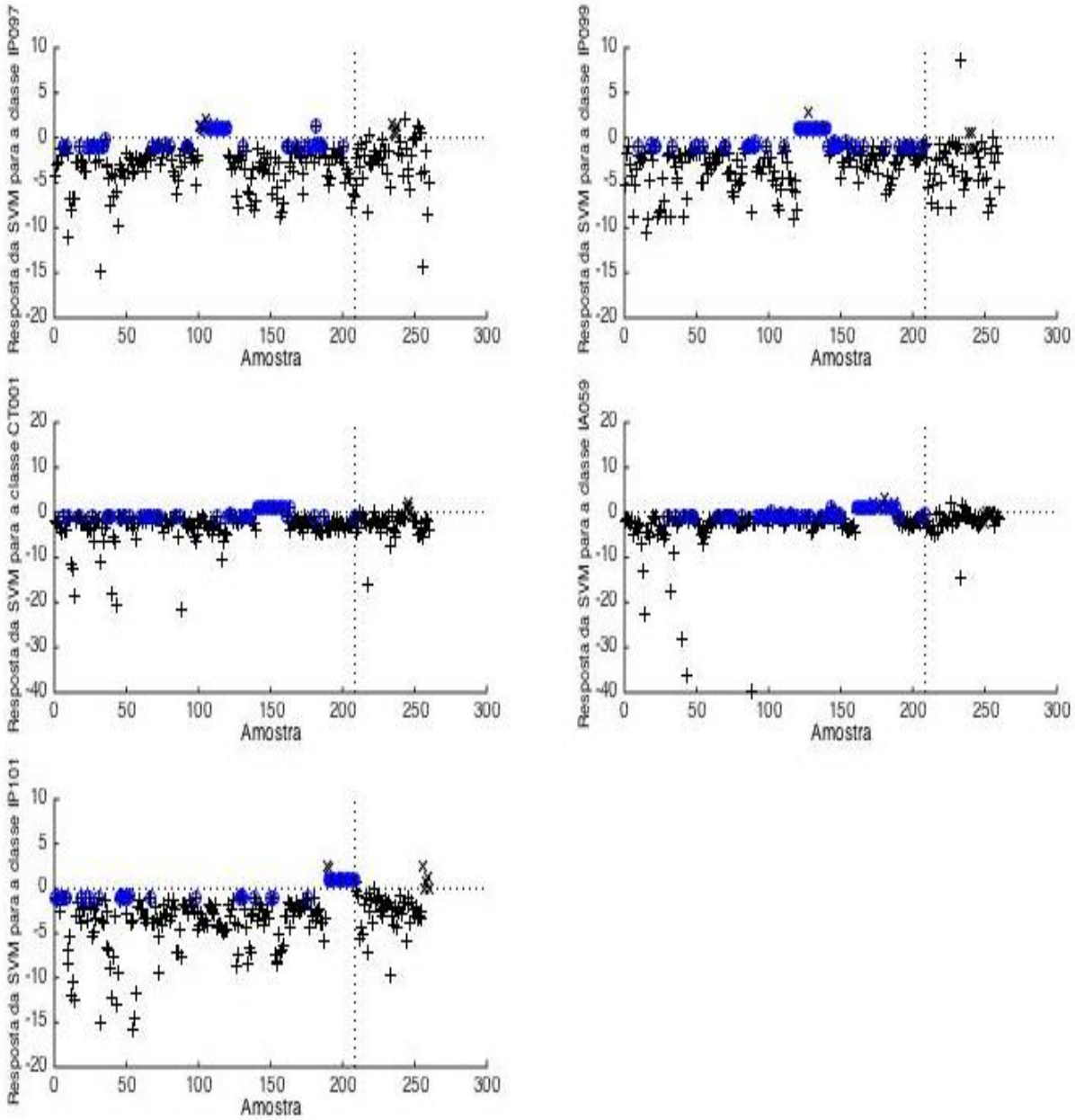


Figura 11. Resposta da melhor SVM. As amostras com sinal “x” são amostras da classe, sinal “+” não pertencem a classe, circuladas em azul são as amostras utilizadas como SVs. A linha vertical tracejada separa as amostras de treinamento das de teste.

Em Lemes (2014) as mesmas amostras utilizadas para a classificação genotípica foram avaliadas através do PLS-DA usando os espectros puros. Nesse modelo foi obtida uma sensibilidade média, para as amostras teste, de 62,9% e uma especificidade média de 98,1%. O mesmo autor utilizou uma rede RBF alimentada com os *scores* do PLS-DA obtidos a partir da primeira derivada dos espectros. Nesse trabalho foi obtida uma sensibilidade média, para as amostras de teste, de 91,4% e uma

especificidade média de 96,28%. A SVM desenvolvida teve um desempenho melhor do que o PLS-DA, porém não foi comparável ao modelo de dois estágios PLS-DA/RBF. Portanto, sugere-se que outras abordagens multiclasse sejam testadas na tentativa de melhorar o desempenho da SVM para a classificação por genótipo.

Na comparação entre os resultados obtidos entre as SVMs, para o problema de classificação geográfica ficou evidente que a utilização das derivadas piorou o desempenho dos modelos. Nas derivadas possíveis bandas importantes que podem estar ocultas são destacadas, porém o mesmo é válido para os ruídos. Já para a classificação genotípica a primeira derivada foi mais eficiente, ou seja, esse tratamento enfatizou a diferença entre as amostras. Para o mesmo problema a segunda derivada não foi eficaz por ter amortizado excessivamente o sinal do FTIR (DONATO *et al*., 2010)

Pode-se observar, também, que a quantidade média de SVs para classificação genotípica (45,8) é maior quando comparado à classificação geográfica (33,5). Este fato evidencia a complexidade do problema de classificação por genótipo já que quanto maior o número de SVs maior será a complexidade do modelo (HAYKIN, 2002).

5. CONCLUSÃO

De acordo com as SVMs desenvolvidas, a classificação geográfica demonstrou-se mais simples que a genotípica. Os resultados obtidos para classificação geográfica foram satisfatórios, porém para a classificação genotípica outras estratégias multiclasse serão testadas na tentativa de melhorar o desempenho dos modelos. Com o desempenho verificado do modelo proposto compreende-se a complexidade de informações do café.

6. REFERÊNCIAS

ABIC. **Produção Mundial de Café – Principais Países Produtores 2014.** Disponível em: <http://www.abic.com.br/publique/cgi/cgilua.exe/sys/start.htm?sid=48> Acesso em: 20 de maio de 2015.

BASSBASI, M.; PLATIKANOV, S.; TAULER, R.; OUSSAMA, A. FTIR-ATR determination of solid non fat (SNF) in raw milk using PLS and SVM chemometric methods. **Food Chemistry,** v. 146, n. 0, p. 250-254, 2014.

BISHOP, C. M. **Neural networks for pattern recognition.** New York, US: Oxford University, 482 p. 2002.

CESARINO, I.; MAZZAFERA, P. Chapter 7 - **Botanical Aspects of the Antioxidant System in Coffee**. In: Preedy, V. R. (Ed.). Coffee in Health and Disease Prevention. San Diego: Academic Press, 2015. p.53-60.

CHANG, C. C.; LIN, C. J. 2011. LIBSVM: A library for support vector machines. **ACM Trans. Intell. Syst. Technol**. 2, 3, Article 27, 2011.

DEVOS, O.; DOWNEY, G.; DUPONCHEL, L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. **Food Chemistry,** v. 148, n. 0, p. 124-130, 2014.

DONATO, E.M.; CANEDO, N.A.P.; ADAMS, A.I.H.; FRÖEHLICH, P.E.; BERGOLD, A.M. Espectrofotometria derivada: uma contribuição prática para o desenvolvimento de métodos. **Revista de Ciências Farmacêuticas Básica Aplicada,** v. 1, n. 2, p. 125-130, 2010.

FERRAO, M. F. et al . **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, São Paulo , v. 30, n. 4, p. 852-859, Aug. 2007

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2ª edição. Porto Alegre: Bookman, 900p. 2001.

ICO. **Relatório sobre o mercado de café – Abril de 2015**. Disponível em : <http://www.agricultura.gov.br/arq_editor/file/15288_relatorio_do_mercado_cafeeiro_-_abril_2015.pdf>. Acesso em: 20 de maio de 2015.

KEMSLEY, E. Katherine; RUAULT, S.; WILSON, Reginald H. Discrimination between Coffea arabica and Coffea canephora variant robusta beans using infrared spectroscopy. **Food Chemistry**. v.54, n.3, p. 321-326, 1995.

KIM, S. S.; PARK, R. Y.; JEON, H. J.; KWON, Y. S.; CHUN, W. Neuroprotective effects of 3,5-dicaffeoylquinic acid on hydrogen peroxide-induced cell death in SH-SY5Y cells. **Phytotherapy Research,** v. 19, n. 3, p. 243-245, 2005.

KLEINWÄCHTER, M.; BYTOF, G.; SELMAR, D. Chapter 9 - **Coffee Beans and Processing**. In: Preedy, V. R. (Ed.). Coffee in Health and Disease Prevention. San Diego: Academic Press, 2015. p.73-81

LEMES, A. L. G. **Aplicação de modelos de dois estágios em problemas de classificação de alta complexidade: segmentação geográfica e genotípica de café arábica.** 2014. 59 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2014.

LI, H.; LIANG, Y.; XU, Q. Support vector machines and its applications in chemistry. **Chemometrics and Intelligent Laboratory Systems,** v. 95, n. 2, p. 188-198, 2009.

LIMA, C. A. M. **Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores-Suporte**, Tese de Doutorado, FEEC/Unicamp, 2004.

LINK, J. V. et al. Geographical and genotypic classification of arabica coffee using Fourier transform infrared spectroscopy and radial-basis function networks**. Chemometrics and Intelligent Laboratory Systems,** v. 135, n. 0, p. 150-156, 2014.

MACHADO, S. R.; PARISE, E. R.; DE CARVALHO, L. Coffee has hepatoprotective benefits in Brazilian patients with chronic hepatitis C even in lower daily consumption than in American and European populations. **The Brazilian Journal of Infectious Diseases,** v. 18, n. 2, p. 170-176, 2014.

MAPA. **Ministério da Agricultura, Pecuária e Abastecimento - 2015**. Disponível em: <http://www.agricultura.gov.br/vegetal/culturas/cafe/saiba-mais>. Acesso em: 20 de maio de 2015.

MARETTO, D. A. **Aplicação de máquinas de vetores de suporte para desenvolvimento de modelos de classificação e calibração multivariada em espectroscopia no infravermelho**. Tese – Instituto de Química. Universidade Estadual de Campinas, 2011.

MATLAB R2014b, The Math Works Inc., USA.

PARREIRA, T. F. **Utilização de Métodos Quimiométricos em Dados de Natureza Multivariada**. Dissertação – Instituto de Química, Universidade de Campinas. Campinas, 2003.

QIU, S.; WANG, J.; GAO, L. Discrimination and Characterization of Strawberry Juice Based on Electronic Nose and Tongue: Comparison of Different Juice Processing Approaches by LDA, PLSR, RF, and SVM. **Journal of Agricultural and Food Chemistry,** v. 62, n. 27, p. 6426-6434, 2014.

RIFKIN, R.; KLAUTAU, A. In Defense of One-Vs-All Classification. **The Journal of Machine Learning Research**, 5, 101–141, 2004.

SAVITZKY, A; GOLAY, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. **Analytical Chemistry***,* 38, p.1627-1639, 1964.

SILVERSTEIN, R. M.; WEBSTER, F. X.; KIEMLE, D. J. **Identificação espectrométrica de compostos orgânicos.** 7. ed. Rio de Janeiro, RJ: LTC, xiv, 490 p., 2007

WANG, N.; LIM, L. T. Fourier Transform Infrared and Physicochemical Analyses of Roasted Coffee. **Journal of Agricultural and Food Chemistry,** v. 60, n. 21, p. 5446-5453, 2012.