UTILIZAÇÃO DE MÁQUINAS DE VETORES SUPORTE PARA CLASSIFICAÇÃO DE BIODIESEL USANDO DADOS OBTIDOS POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO.

Werickson Fortunato de Carvalho Rocha [1], *Luiz Henrique da Conceição Leal* [2], *Gabriel Fonseca Sarmanho* [3]

[1] Inmetro, Rio de Janeiro, Brasil, wfrocha@inmetro.gov.br
[2] Inmetro, Rio de Janeiro, Brasil, lhleal@inmetro.gov.br
[3] Inmetro, Rio de Janeiro, Brasil, gfsarmanho@inmetro.gov.br

Resumo: O trabalho tem por objetivo apresentar uma aplicação da técnica de aprendizado supervisionado denominada Máquinas de Vetores Suporte (Support Vector Machines - SVM), ferramenta utilizada para classificação de padrões. Utilizando a técnica do infravermelho próximo, foram coletadas 50 amostras de cada um dos três tipos diferentes de biodiesel analisados, conforme a sua matéria prima (óleo vegetal): amendoim, canola e milho. Pela relevância das informações contidas, foi selecionada a faixa spectral que compreende o comprimento de onda indo de 3300 a 670 (cm^{-1}). Dentre as possíveis funções Kernel foi selecionada a Polinomial. Os resultados mostraram que a técnica SVM é uma ferramenta promissora para o reconhecimento de padrão dos tipos de biodiesel, já que somente para um dos tipos de óleo vegetal (canola) houve erro de classificação das amostras. Tanto para milho, quanto para o amendoim todas as amostras foram corretamente classificadas.

Palavras chave: Máquinas de Vetores Suporte, Infravermelho Próximo, Biodiesel.

1. INTRODUÇÃO

Segundo Lima e colaboradores [1] a maior parte de toda a energia consumida no mundo provém do petróleo, carvão e do gás natural. Entretanto, essas são fontes não renováveis e possuem previsão de esgotamento em um futuro próximo [2]. Além disso, os combustíveis fósseis são muito poluidores afetando o meio ambiente de forma bastante agressiva, o que faz a população mundial buscar soluções para tais problemas. Os óleos vegetais, como alternativa de combustível, começaram a ser estudados no final do século XIX por R. Diesel, sendo que estes eram usados *in natura*, ou seja, na forma de óleo. Mas o uso direto nos motores apresenta muitos problemas, como por ex.: acúmulo de material oleoso nos bicos de injeção, a queima do óleo é incompleta, forma depósitos de carvão na câmara de combustão, o rendimento de potência é baixo e, como resultado da queima, libera a acroleína (propenal) que é tóxica [3]. Porém, alternativas têm sido consideradas para melhorar o desempenho de óleos vegetais em motores do ciclo diesel, como por ex., diluição, micro-emulsão com metanol ou etanol, craqueamento catalítico e reação de transesterificação com álcoois de cadeia pequena. Dentre essas alternativas, a reação de transesterificação tem sido a mais usada, visto que o processo é relativamente simples e o produto obtido (biodiesel) possui propriedades muito similares às do petrodiesel [3].

O biodiesel é um combustível que pode se obtido a partir de matérias-primas derivadas de óleos vegetais, tais como canola, palma, girassol e amendoim. Essa possibilidade do uso de diferentes fontes de óleos vegetais na produção de biodiesel gera problemas relacionados à produção e qualidade deste combustível, abrindo precedentes para possíveis fraudes fiscais.

Nesse contexto, a metrologia química, juntamente com a Quimiometria, tem como intuito evitar a ocorrência dessas fraudes, assegurando a qualidade, comprovando a eficiência e demonstrando a exatidão dos resultados de medições, segundo as normas estabelecidas pela Agência Nacional do Petróleo (ANP).

Dessa forma, a calibração multivariada pode ser empregada para determinação de biodiesel utilizando diversos modelos matemáticos como, por exemplo, modelos de calibração linear clássica, modelos de calibração linear robusta e até mesmo modelos de calibração não-linear. Esses métodos de calibração fazem uso de um vetor de medidas instrumentais para cada amostra possibilitando análises mesmo na presença de interferentes, desde que esses interferentes estejam presentes nas amostras de calibração (vantagem de primeira ordem), determinações simultâneas e análises sem resolução.

2. MÁQUINAS DE VETORES SUPORTE

As Máquinas de Vetores Suporte (Support Vector Machines – SVM) constituem, de acordo com a nomenclatura de aprendizado em máquina (Machine Learning), uma técnica de aprendizado supervisionado. Esta metodologia é utilizada como ferramenta de calibração não-linear para reconhecimento de padrões (classificação e análise de regressão).

Formalmente, o método consiste em minimizar o risco estrutural, isto é, a probabilidade de classificar erroneamente um dado previamente desconhecido selecionado

aleatoriamente de uma fixada, mas desconhecida, distribuição de probabilidade [4].

Analiticamente, consiste em encontrar um α que minimize a função objetivo:

$$Q(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i; x_j) \quad (1)$$

Sujeita as restrições:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \qquad \forall i \end{cases} \quad (2)$$

Sendo este um problema de programação quadrática em sua forma dual pode-se também maximizar a seguinte função objetivo (sujeita as mesmas restrições):

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i; x_j) \quad (3)$$

Para a tarefa de classificação a função de decisão é da forma:

$$y = sign(\sum_{i=1}^{n} y_i \alpha_i K(x; x_i) + b) \quad (4)$$

em que $x \in \mathbb{R}^d$ é o vetor de input d-dimensional da amostra de treinamento, $y \in \{-1; 1\}$ é o rótulo da classe, x_i é a i-ésima amostra de treinamento, y_i é o rótulo da classe da i-ésima amostra de treinamento, *n* é o número de amostras de treinamento, $K(x; x_i)$ é a função de Kernel (Tabela 1), $\alpha = \{\alpha_1, \ldots, \alpha_n\}$ (multiplicadores de Lagrange) e b são parâmetros do modelo [5-6].

Tabela 1. Parâmetros da função Kernel

Kernel	Fórmula	Parâmetros
Linear	$K(x_i; x_j) = x_i^{\mid} \cdot x_j$	-
Polinomial	$K(x_i; x_j) = \left(\gamma x_i^{\mid} \cdot x_j + r\right)^d$	γ , r, d
Radial Basis Function (RBF)	$K(x_i; x_j) = exp\left(-\gamma \lVert x_i - x_j \rVert^2\right)$	γ
Sigmoidal	$K(x_i; x_j) = tanh\left(\gamma\left(x_i^{\mid} \cdot x_j\right) + r\right)$	γ, r

Na segunda restrição do problema de programação quadrática há uma constante C que limita superiormente os multiplicadores de Lagrange. Esta é uma constante positiva que determina o "trade-off" entre a maximização da margem, isto é, da distância entre o classificador e a amostra mais próxima de cada classe e a minimização do erro de classificação na amostra de treinamento. A constante C pode ser vista como o custo de classificar erroneamente uma amostra [7].

Tal constante é um parâmetro de penalização do modelo a ser otimizado, sendo esta definida pelo usuário. Caso o valor atribuído a ela seja muito grande, tem-se uma alta penalização para os pontos não separáveis e, dessa forma, pode-se armazenar muitos vetores suporte superestimados. Já quando C é muito pequena os mesmos podem ser subestimados.

As Máquinas de Vetores Suporte podem ser classificadas em dois tipos conforme será apresentado na seção seguinte.

2.1. Tipos de Máquinas de Vetores Suporte

O modelo apresentado anteriormente refere-se à Máquinas de Vetores Suporte para Margens Suaves (Soft-Margin Support Vector Machines) o qual é utilizado quando as amostras não são linearmente separáveis (figura 1). Nos casos em que é possível a separação de forma linear das amostras (figura 2) utiliza-se o modelo de Máquinas de Vetores Suporte para Margens Rígidas (Hard-Margin Support Vector Machines) [7].

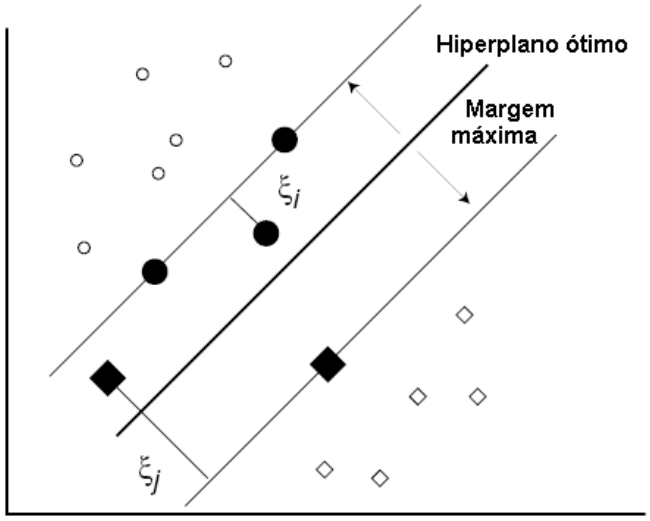

Fig. 1. Máquinas de Vetores Suporte para Margens Suaves.

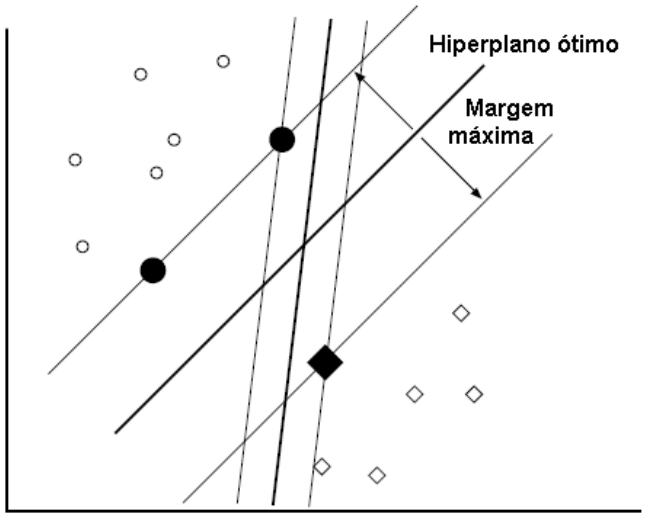

Fig. 2. Máquinas de Vetores Suporte para Margens Rígidas.

A única diferença entre Máquinas de Vetores Suporte para Margens Suaves e Máquinas de Vetores Suporte para Margens Rígidas, considerando a forma dual do problema de programação quadrática, é que α_i (i=1,2,...,n) não pode exceder C [8].

No que concerne a forma primal, as diferenças entre os dois tipos de Máquinas de Vetores suporte são que: há inclusão de uma segunda parcela à direita da função objetivo (a qual deve ser minimizada); e a introdução de variáveis de folga (ξ_i) nas Máquinas de Vetores Suporte para Margens

Suaves (ver equações 5 e 6 abaixo). Assim, o problema de programação quadrática na forma primal é apresentado a seguir:

$$Q(w; b; \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (5)$$

Sujeito as restrições:

$$\begin{cases} y_i\left(w^{\mid}x_i + b\right) \geq 1 - \xi_i \quad \forall i \\ \xi_i \geq 0 \end{cases} \quad (6)$$

em que ξ_i é a i-ésima variável de folga (figura 1) e w é um vetor normal perpendicular ao hiperplano.

2.2. Erro de classificação

O erro de classificação mede a capacidade do modelo adotado de classificar corretamente um objeto, neste caso o biodiesel, pertencente a uma determinada classe [9]. O cálculo do erro é dado pela seguinte fórmula:

$$E = \frac{N_e}{N} \quad (7)$$

em que N_e é o número de amostras classificadas erroneamente e N é o número total de amostras. Na forma de percentagem, pode-se escrever o erro de classificação de acordo com a equação 8:

$$E(\%) = \frac{N_e}{N} \cdot 100 \quad (8)$$

3. MATERIAIS E MÉTODOS

Foram sintetizados três tipos de biodiesel a partir de diferentes matérias-primas: óleo vegetal de amendoim, canola e milho, purificados no Laboratório de Motores, Combustíveis e Lubrificantes (Lamoc) do Inmetro.

Após a síntese, foram feitas 50 amostras de misturas biodiesel/diesel cujas concentrações estão na faixa de 2% (V/V) a 90% (V/V) em recipientes de 5 mL para cada tipo de biodiesel. Todas as amostras foram agitadas durante 1 minuto para garantir a homogeneização. Devido a um erro de medição excluiu-se uma amostra do biodiesel de óleo de canola perfazendo-se um total de 149 amostras. As análises foram realizadas no espectrômetro GX Spectrum da Perkin-Elmer com acessório de refletância difusa, com uma resolução de 4 cm^{-1} com 10 varreduras por espectro.

O programa utilizado para a implementação dos modelos SVM (estimação dos parâmetros e construção do gráfico) foi o "R" utilizando o pacote "e1071" disponível no endereço eletrônico http://cran.fiocruz.br/.

4. RESULTADOS

Na prática há poucos conjuntos de dados em que as amostras possam ser linearmente separáveis [10]. Neste contexto o modelo adotado no presente trabalho foi o de Máquinas de Vetores Suporte para Margens Suaves.

A técnica de Máquinas de Vetores Suporte depende da escolha da função Kernel e dos respectivos parâmetros. Diversas funções Kernel foram testadas e a que apresentou melhores resultados foi a polinomial a qual depende dos parâmetros γ, r e d (tabela 1) bem como do parâmetro de penalização C. O pacote "e1071" apresenta a função "*tune.svm()*" na qual pode-se testar diversas combinações dos parâmetros γ e C com a finalidade de buscar o par de valores que produzem o melhor ajuste. Para maiores detalhamentos sugere-se consultar a referência [11]. Os parâmetros que produziram o melhor ajuste foram $\gamma = 10^{-4}$ e C = 10, sendo assim a segunda restrição (equação 2), do problema de programação quadrática definido na equação 1, é da forma:

$$0 \leq \alpha_i \leq 10, \forall i \quad (9)$$

Neste contexto (adotando-se o parâmetro r = 1) tem-se a seguinte função Kernel polinomial:

$$K(x_i; x_j) = \left(10^{-4}x_i^{\mid} \cdot x_j + 1\right)^d \quad (10)$$

em que *d* representa o grau do polinômio.

A escolha do grau do polinômio foi realizada com auxílio do gráfico apresentado na figura 3. Diversos modelos foram testados com diferentes graus para a função Kernal polinomial e os respectivos erros de classificação foram comparados. A partir do grau 3 a redução no erro de classificação não é significativa, sendo assim o grau adotado para a função Kernel polinomial foi o grau 3.

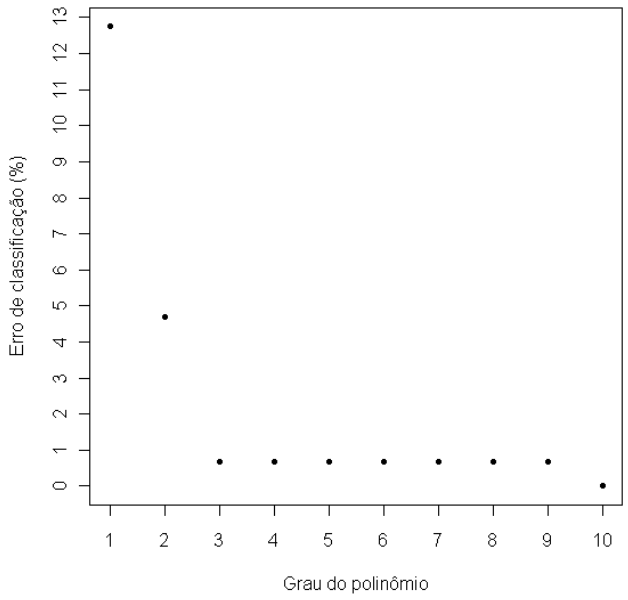

Fig. 3. Erro de classificação versus o grau do polinômio, usando a função Kernel polinomial.

O modelo de Máquina de Vetores Suporte adotado apresentou um erro de classificação de 0,67% (v/v). Na Tabela 2 observa-se que apenas uma dentre todas as amostras foi classificada erroneamente.

Tabela 2. Classificação do biodiesel

Observado \ Classificado	Amendoim	Canola	Milho
Amendoim	50	0	0
Canola	0	48	1
Milho	0	0	50

O biodiesel de óleo de canola foi classificado erroneamente como óleo de milho (tabela 2). É importante ressaltar que, dentre os diversos modelos testados, o óleo de canola foi o que apresentou maior dificuldade de classificação dentre os três tipos de óleo vegetal analisados neste estudo.

5. CONCLUSÃO

A técnica de Máquinas de Vetores Suporte é um método muito eficiente para classificação de biodiesel segundo o tipo de óleo vegetal [9]. No presente trabalho a função Kernel que produziu a melhor classificação foi a polinomial de grau 3, sendo que, dos três tipos de óleo vegetal analisados, o que apresentou maior dificuldade de classificação foi o óleo de canola.

Convém ressaltar que neste trabalho, como foi aplicado um método de reconhecimento de padrão não supervisionado, não foi realizada a classificação de novas amostras de acordo com o modelo estimado.

A busca de uma técnica eficiente de classificação é útil para evitar a ocorrência de fraude na produção de Biodiesel assegurando desse modo a qualidade do produto.

AGRADECIMENTOS

Agradecemos ao pessoal do Lamoc (Laboratório de Motores, Combustíveis e Lubrificantes) do Inmetro pela importante ajuda com a realização d síntese das amostras de biodiesel.

REFERÊNCIAS

[1] J. R. O. Lima, R. B. Silva, C. C. M. Silva, L. S. S. Santos, J. R. Santos Jr., E. M. Moura, C. V. R. Moura, "*Biodiesel de babaçu (Orbignya sp.) obtido por via etanólica*", Química Nova, v. 30, n 3, pp 600-603, Piauí, Brasil, Junho, 2007.

[2] U. Schuchardt, R. Sercheli, R. M. Vargas, "*Transesterification of Vegetable Oils: a Review*", Journal of the Brazilian Chemical Society, v9, n1, pp 199-210, São Paulo, Brazil, January, 1998.

[3] R. Gardner, S. Kazi, E. M. Ellis, "*Detoxication of the environmental pollutant acrolein by a rat liver aldo-keto reductase*", Toxicology Letters, v. 148, pp 65-72, Scotland, UK, March, 2004.

[4] L. Walawalkar, M. Yeasin, A. M. Narasimhamurthy, R. Sharma, "*Support Vector Learning for Gender Classification Using Audio and Visual Cues: A Comparison*", First International Workshop SVM, pp 144-159, Niagara Falls, Canadá, August, 2002.

[5] C.-W. Hsu, C.-C. Chang, C.-J. Lin, "*A practical guide to support vector classification*", Technical report, Department of Computer Science, National Taiwan University, Taiwan, April, 2010.

[6] R. Collobert, Y. Bengio, S. Bengio, "*Scaling Large Learning Problems with Hard Parallel Mixtures*", First International Workshop SVM, pp 8-23, Niagara Falls, Canadá, August, 2002.

[7] Y. Ma, X. Ding, "*Face Detection Based on Cost-Sensitive Support Vector Machines*", First International Workshop SVM, pp 8-23, Niagara Falls, Canadá, August, 2002.

[8] S. Abe, *Support Vector Machines for Pattern Clasification*. 2 ed. New York, Springer-Verlag, 2009.

[9] R. M. Balabin, R. Z. Safieva, "*Biodiesel Classification by base stock type (vegetable oil) using near infrared spectroscopy data*", Analytica Chimica Acta, pp 190-197, Russia, January, 2011.

[10] L. Hamel, *Knowledge Discovery with Support Vector Machines*, New Jersey, Wiley, 2009.

[11] A. Karatzoglou, D. Meyer, K. Hornik, "*Support Vector Machines in R*", Journal of Statistical Software, v. 15, n. 9, pp 1-28, Austria, April, 2006.