SIC 2018 SEMINÁRIO DE INICIAÇÃO CIENTÍFICA
21 SET 2018
PROPIT Pró-Reitoria de Pós-Graduação, Pesquisa e Inovação Tecnológica
sic.unifesspa.edu.br

# Desenvolvimento e melhoramento de estratégias de seleção de variáveis em problemas de classificação: Otimização por enxame de partículas- PSO

Weverton D Lucas Moura Marinho
Adriano de Araújo Gomes

**Agência financiadora:** CNPq

**Resumo:** Neste trabalho foi desenvolvido uma nova estratégia de seleção de variáveis em problemas de classificação baseado em otimização por enxame de partículas- PSO. A presente proposta foi avaliada em um estudo de caso simulado e na identificação de qualidade de óleo lubrificante de motor ciclo diesel usando espectrometria no infravermelho no modo ATR. Foram adquiridos 24 frascos de 1 L de óleo lubrificantes de distintos lotes, os mesmos foram adquiridos junto a empresa Chevron Brasil, realizou-se ensaios de degradação simulada por aquecimento e exposição à radiação UV a distintos tempos. Os espectros de infravermelho no modo ATR foram registrados em um equipamento AGILENT CARY 630 FTIR na faixa de 650 a 4000 cm-1, com resolução de 2 $cm^{-1}$. Os modelos de classificação foram construídos em ambiente MatLab, os resultados obtidos permitiram a discriminação das amostras de óleo lubrificante em conformidade de não conformes, informando que o método proposto é eficaz no controle de qualidade de óleo lubrificante e que o método PSO é promissor como ferramenta de seleção de variáveis em problemas de classificação.

**Palavras chave:** PSO-LDA; infravermelho; óleo lubrificante.

## 1. INTRODUÇÃO

Ferramentas quimiométricas são muito úteis no que diz respeito a otimização de condições experimental e compilação de dados químicos para tomada de decisões (ANTONIJE, 2016). Aplicações da quimiometria vem sendo reportada nas diversas áreas da química e Ciências correlatas. Contudo, para uma maior difusão destas ferramentas é preciso o desenvolvimento de software amigáveis e de fácil operação. Técnicas instrumentais, tais como espectroscópicas e imagens digitais em combinação com modelos multivariados tem permitido o desenvolvimento de novas metodologias analíticas que portam uma série de vantagens como: rapidez, baixo custo, robustez, determinações simultâneas e menores consumo de amostras e reagentes. Contudo a qualidade final de um modelo quimiometrico está ligada diretamente com as variáveis de entrada (PESSOA, 2015). As técnicas instrumentais modernas são conhecidas por gerar uma grande quantidade de informação por amostra em um pequeno intervalo de tempo (GOMES, 2015). Entretanto parte da informação gerada não é informativa e/ou redundante, o que faz necessário o uso de estratégias de seleção de variáveis como etapa previa na construção dos modelos.

Inspirado em estudos no campo da neurociências, psicologia cognitiva e ciências comportamentais, o conceito de inteligência dos enxames (SI) foi introduzido no domínio da computação e inteligência artificial em 1989 (SHAMSIPUR, 2007), os algoritmos baseados em enxames surgiram como uma poderosa família de técnicas de otimização, inspiradas no comportamento sociais coletivo de animais (TABAKHI, 2015). No contexto de seleção de variáveis, empregando uma abordagem via PSO, o enxame de partida é inicializado randomicamente. O enxame inicial consiste em possíveis soluções para o problema, ou seja, são subconjuntos de variáveis. A cada partícula do enxame está associada uma posição ($x_i$). A cada iteração do método PSO, posição e velocidades são atualizados até que seja encontrado o ótimo global.

## 2. MATERIAL E MÉTODOS

### 2.1 Método proposto

O algoritmo PSO acoplado à LDA proposto neste trabalho é composto de cinco etapas: codificação binaria em que foi atribuído 1 para variáveis incluídas na construção do modelo e 0 para não incluídas, inicialização em modo randômico, avaliação tendo como base a função de custo (*Gcost*) que informa o risco médio do modelo LDA, cometer um erro de classificação tendo como base um subconjunto de variáveis, atualização das posições e velocidades das partículas do enxame e a etapa final é a atualização do enxame.

### 2.2 Estudo de caso com dados simulados

Os dados foram simulados com intuito de mimetizar um típico problema de classificação com duas classes alvo definidas. Foi empregue somas de perfis gaussianos para obter espectros simulados. Para avaliar a classificação, foram gerados espectros simulados de 90 amostras pertencentes a duas classes (Classe 1 com 45 e amostras e classe 2 com 45 amostras). Estas amostras foram particionadas em Treinamento (25 amostras por classe), validação (10 amostras por classe) e teste (10 amostras por classe) empregando métodos Kernnard-Stone (KS).

### 2.3 Estudo de casos com dados reais de óleo lubrificante de motor de ciclo diesel

Amostra de óleo lubrificantes (24 frascos de 1 L) de distintos lotes foram adquiridos junto a empresa Chevron Brasil. De cada amostra foram coletadas três alíquotas para realização dos ensaios de degradação simulada por aquecimento e exposição à radiação UV a distintos tempos (de 0,5 a 6 horas). Os espectros de infravermelho no modo ATR foram registrados em um equipamento AGILENT CARY 630 FTIR na faixa de 650 a 4000 $cm^{-1}$, com resolução de 2 $cm^{-1}$.

## 3. RESULTADOS E DISCUSSÃO

O método proposto foi implementado como atualização da interface: **LINEAR DISCRIMINANT ANALYSISVARIABLE SELECTION TOOLBOX** desenvolvida em ambiente MatLab® (2010b) para construção de modelos LDA com ou sem associação a métodos de seleção de variáveis. O algoritmo desenvolvido neste trabalho, denominado do PSO-LDA, assim como todos os cálculos envolvendo os dados simulados foram realizados em ambiente MatLab 2010a, cuja licença foi cedida pelo grupo parceiro LAQA/UFPB.

### 3.1 Estudo de caso com dados simulados

Na Figura 1.A ilustra os espectros simulados para o estudo de casos envolvendo classificação multivariada LDA, pode-se observar uma forte sobreposição dos espectros das classes 1 e 2. Porém, este problema de classificação deve ser solucionado por métodos de reconhecimento de padrões supervisionado, como o método de análise de componentes principais (PCA). A Figura 1.B, mostra o gráfico de escores para (PCA) deste problema, note que há que informações nestes dados que permite a discriminar a classe 1 e 2 que foram acessados pelo modelo (PCA).

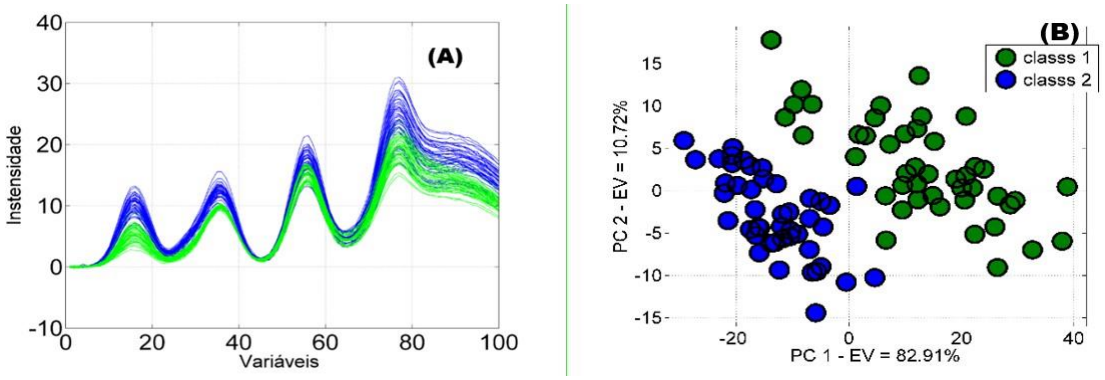


**Figura 1**: Em (A) Espectros simulados usados. Linhas azuis classe 1 e verdes classe 2. Em (B) Gráfico de escores.

Após analisar os dados por (PCA) foram construídos modelos baseados em analise discriminante por mínimos quadrados parciais (PLS-DA), seleção de variáveis baseados no algoritmo genético (GA-LDA) e o método proposto (PSO-LDA), afim de analisar seu poder discriminante. Os resultados são mostrados na Tabela 1, vemos que todos os modelos foram hábeis para produzir resultados altamente seletivos e específicos. Mostrando que o método proposto é tão eficaz quanto a abordagens já consolidadas na literatura como PLS-DA e o GA-LDA.

| Método | Treinamento | Validação | Teste |
|---|---|---|---|
| | Taxa de acerto | Taxa de acerto | Taxa de acerto |
| *PLS-DA (6) ** | 100% | 100% | 100% |
| *GA-LDA (1)* | 96.6% | 96.6% | 100% |
| *PSO-LDA (3)* | 100% | 100% | 100% |

**Tabela 1:** resumo do ajuste e predição.

### 3.2 Estudo de casos com dados reais de óleo lubrificante de motor de ciclo diesel

Na Figura 2A são apresentados os perfis dos espectros brutos das amostras de óleo lubrificante, é possível observar que ocorre uma forte sobreposição dos perfis, não sendo possível distinguir amostras em conformidades e não conformes por UV e aquecimento tendo como base uma inspeção visual dos espectros, construísse então o modelo PLS-LDA afim de discrimina as amostras de óleo lubrificante. As variáveis selecionadas pelo PLS-LDA estão mostradas na Figura 2B. Os modelos LDA resultante classificou corretamente todas amostras como pode ser visualizado no gráfico de dispersão (DF1 versus DF2) dos escores de Fisher.

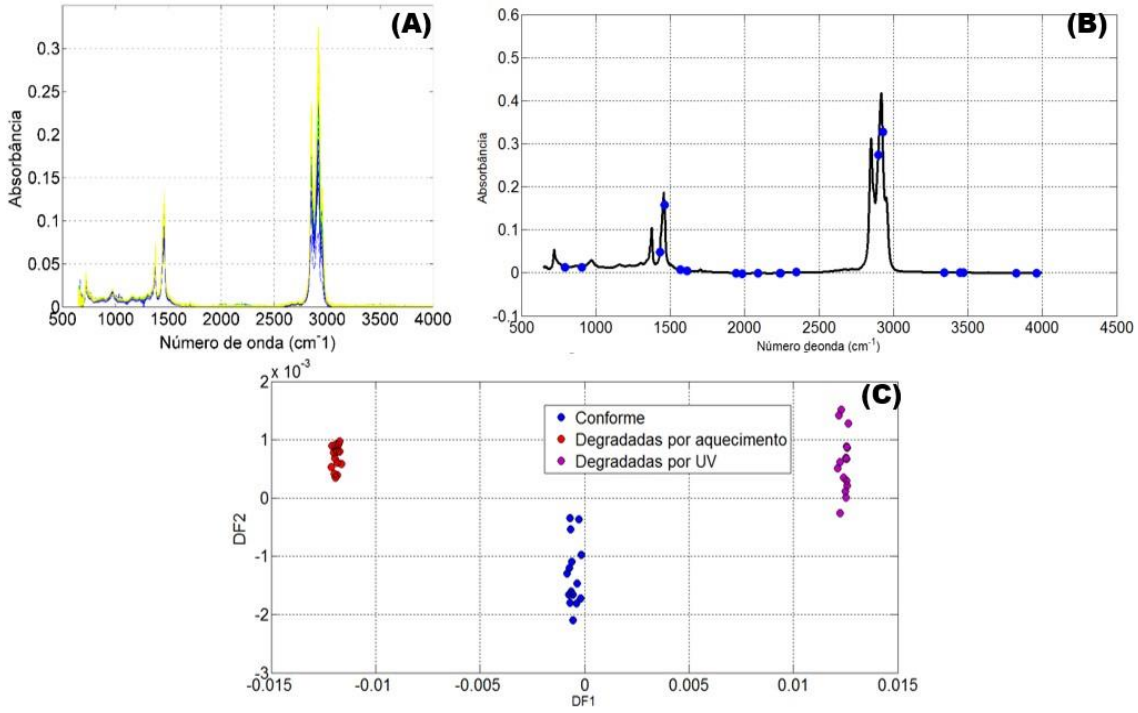


**Figura 2**: Em (A) espectros ATR-FTIR (azul, óleo não degradado, verde: degradação por aquecimento e amarelo degradação por UV). Em (B) variáveis selecionadas PSO-LD. Em (C) Gráfico dos escores de Fisher baseados nas variáveis selecionas pelo PSO-LDA.

Comparou-se o método proposto com modelos baseados em analise discriminante já consolidado na literatura como, o dos mínimos quadrados parciais (PLS-DA), seleção de variáveis

baseados no algoritmo genético (GA-LDA). Os resultados são mostrados na Tabela 2. Como pode ser observar que os melhores resultados foram obtidos para modelos baseados em seleção de variáveis, especialmente para o método proposto, que alcançou 100% de classificação correta.

**Tabela 2:** resumo do ajuste e predição.

| Método | Treinamento | Validação | Teste |
|---|---|---|---|
| | Taxa de acerto | Taxa de acerto | Taxa de acerto |
| *PLS-DA (6)** | 90% | 92% | 71% |
| *GA-LDA (9)* | 93% | 94% | 98% |
| *PSO-LDA (3)* | 100% | 100% | 100% |

## 4. CONCLUSÃO

Como pode ser observado, os resultados obtidos sugerem que a metodologia proposta (PSOLDA) se mostrou eficaz podendo ser considerado uma alternativa útil na solução de problemas envolvendo reconhecimento de padrões em dados químicos. Os espectros ART-FTIR portam a informação química necessária para discriminar amostras de óleo lubrificante em conformidade e de amostras em não conformidade.

## 5. REFERÊNCIAS E CITAÇÕES


ANTONIJE, E. Onjia. **Chemometric Approach to the Experiment Optimization and Data Evaluation in Analytical Chemistry.** Ed. Única. Faculty of Technology and Metallurgy, University of Belgrade Karnegijeva: Belgrade, 2016.

GOMES, Adriano de Araújo. **Algoritmo das Projeções Sucessivas para Seleção de Variáveis em Calibração de Segunda Ordem.**pg.126. Tese de Doutorado - Programa De Pós-Graduação Em Química- Centro De Ciências Exatas E Da Natureza Departamento De Química, Universidade Federal Da Paraíba, Paraíba 2015.

PESSOA, Carolina de Marco. **Aperfeiçoamento do Algoritmo de Colônia de Formigas para o Desenvolvimento de Modelo Quimiometricos.** 2015. pg. 88. Dissertação de Mestrado- Programa de Pós-Graduação em Engenharia Química – Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, Porto Alegre 2015.

SHAMSIPUR, Mojtaba; ZARE-SHAHABADI, Vali; HEMMATEENEJAD, Bahram; AKHOND, Morteza. Ant colony optimisation: Ant colony optimisation: a powerful tool for wavelength selection. **Journal of Chemometrics.** Vol. 20. pg. 398-405, Jan 2007.

TABAKHI, S.; MORADI, P. Relevance-redundancy feature selection based on ant colony optimization. **Pattern Recognition.** Vol.48.pg. 2798-2811, Abr 2015.