Algoritmo Evolutivo Multi-objetivo de Tabelas para Seleção de Variáveis em Calibração Multivariada

Carlos Antônio Campos Jorge
Telma Woerle de Lima
Lucas de Almeida Ribeiro
Universidade Federal de Goiás
Goiânia, Goiás
Email: telma@inf.ufg.br

Arlindo R. G. Filho
Instituto Tecnológico da Aeronaútica
São José dos Campos, São Paulo
Email: arlindo@ita.br

Clarimar José Coelho
Pontifícia Universidade
Católica de Goiás
Goiânia, Goiás

Anderson da Silva Soares
Alexandre C. B. Delbem
Instituto de Ciências Matemáticas
e de Computação - USP
São Carlos, São Paulo

***Resumo*—This paper proposes a new multiobjective evolutionary algorithm. In particular, this algorithm makes use of a structure of subsets stored in a data structure called table in which the best individuals from each target are considered preserved. This approach is compared in this paper with traditional evolutionary algorithms monoobjective. A case study is presented a problem multivariate calibration which involves the prediction of protein concentration in samples of whole wheat from spectrophotometric measurements wave. The results show that the proposed formulation has a smaller prediction error as compared to the formulation monoobjective. Additionally, it was observed that the templates are obtained from the proposed formulation using a lower number of variables. Finally, a study of sensitivity to noise in the models obtained by multi-objective formulation showed better result.**

I. Introdução

Em diversos problemas das áreas médica, biológica, industrial, química entre outras, é de grande interesse verificar se duas ou mais variáveis estão relacionadas de alguma forma. Dentre essas áreas, destaca-se a quimiometria, que é uma área que se refere à aplicação de métodos estatísticos e matemáticos a dados de origem química. Em geral, as concentrações dos compostos de interesse numa amostra são determinadas através da resolução de um sistema de equações simultâneas obtidas pela lei de Beer. Curvas de calibração são construídas para cada comprimento de onda a partir de concentrações de referência de cada analito a fim de estabelecer a relação entre concentração e intensidade de absorção [1].

Para expressar a relação entre as variáveis pode-se estabelecer um modelo matemático que relacione as variáveis dependentes com as independentes [2] . Nesse contexto, a regressão linear múltipla, uma técnica estatística utilizada para construir os modelos que descrevem essas relações, pode ser aplicada.

Nas últimas décadas, dentro da área de quimiometria, houve um aumento significativo no uso de computadores para analisar dados químicos, em parte devido aos recentes avanços em hardware de aquisição de dados e do interfaceamento de instrumentos aos computadores. Tais avanços proporcionaram um aumento substancial a quantidade de informação, muitas vezes complexa e variada. Neste contexto, o uso de algoritmos de seleção de variáveis é importante para escolher um subconjunto de variáveis que descreva adequadamente o problema de predição de um determinado analito.

A literatura sobre o problema de seleção de variáveis no contexto da quimiometria (veja referencias [3]–[6]) indica que o algoritmo evolutivo possui duas limitações para o problema analisado: (i) seleciona um número de variáveis maior do que modelos clássicos como o algoritmo das projeções sucessivas (APS) e (ii) os modelos gerados possuem uma maior sensibilidade à presença de ruídos instrumentais. Recentemente, Lucena [7], com o objetivo de reduzir o problema (i), propôs uma formulação multi-objetivo utilizando o algoritmo NSGA-II que minimiza o erro de predição e o número de variáveis selecionadas de forma simultânea. Os resultados demonstraram que a formulação com dois objetivos conseguiu minimizar simultaneamente os objetivos analisados a níveis melhores do que os algoritmos tradicionais. Entretanto, o problema de sensibilidade à ruído não foi considerado.

A formulação original do algoritmo NSGA-II utilizada por Lucena et. al [7] foi proposta por Deb et. al [8]. Neste trabalho foi proposto o uso do conceito de não-dominância, em que não é possível inferir se um determinado indivíduo é melhor do que o outro (não-dominado) em todos os objetivos considerados. Apesar de apresentar uma técnica eficiente em termos de otimização multi-objetivo com algoritmos genéticos, o autor afirma que tal algoritmo pode não ser eficaz para problemas em que três ou mais objetivos são considerados.

Em [9] foi proposto a análise multi-objetivo para seleção de variáveis envolvendo a capacidade de predição e a sensibilidade à ruído. Verificou-se que a associação de tal objetivo resultou em modelos mais robustos para determinadas configurações das amostras para construção do modelo. Porém os resultados obtidos com os objetivos se restringiram somente a análise do modelo já construído. Uma abordagem não coberta no trabalho é a possibilidade de incorporar tais objetivos na seleção de variáveis no intuito de construir modelos com baixa sensibilidade.

Neste cenário, este trabalho propõe um algoritmo evolutivo multi-objetivo baseado em uma nova forma de seleção de indivíduos. Em particular faz-se o uso de uma seleção de subpopulações em estrutura de dados denominada tabelas. Espera-se que o algoritmo seja capaz de gerenciar mais de dois objetivos de forma a encontrar soluções satisfatórias para o problema de seleção de variáveis. Tal proposta visa preservar os melhores indivíduos para os objetivos considerados: (i) o erro de predição, (ii) o número de variáveis, (iii) a sensibilidade

a ruído e (iv) função de agregação para ponderação das anteriores. O estudo de caso da calibração multivariada envolve a determinação do teor de proteína em um conjunto de amostras de trigo de dados obtidas através de espectrofotometria próxima ao infravermelho.

II. Calibração Multivariada e Seleção de Variáveis

Análise química quantitativa é a ciência da determinação da concentração de um ou mais substancias presentes em uma amostra. O estado da arte dessa aplicação é o uso de técnicas espectrofotométricas que medem a interação entre o objeto em análise e a energia radiada [10].

Essa interação é baseada pela Lei de Lambert-Beer [11], ilustrada na Figura 1, na qual mostra uma amostra recebendo uma radiação P_0 , e atravessando com uma energia menor P. A energia absorvida da amostra pode ser medida com um espectrofotômetro e relacionado com a concentração de propriedade [12]. Portanto, a intensidade de absorbância é numericamente dada por

$$x(\lambda) = \log \frac{P_0(\lambda)}{P(\lambda)} \tag{1}$$

onde $P_0(\lambda)$ é a radiação emitida pelo equipamento e $P(\lambda)$ é a radiação emitida pela amostra no comprimento de onda λ.

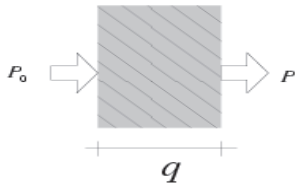


Figura 1. Processo de espectroscopia de absorção.

Para obter a concentração da amostra, é necessário irradiar diferentes comprimentos de onda simultaneamente. Neste cenário, é normal comprimentos de onda se sobreporem e, consequentemente, dois ou mais sinais enviarem a mesma informação. Em termos algébricos as ondas sobrepostas significam alta correlação entre as variáveis e podem induzir a problemas matemáticos no processo de regressão [13].

Seja uma amostra incluindo duas absorbância (A and B) com com sobreposição espectral $\lambda(1)$ e $\lambda(2)$, é possível obter y_A and y_B tal que

$$\begin{aligned} x(\lambda_1) &= k_A(\lambda_1)y_A + k_B(\lambda_1)y_B \\ x(\lambda_2) &= k_A(\lambda_2)y_A + k_B(\lambda_2)y_B \end{aligned} \tag{2}$$

$$\begin{aligned} \begin{bmatrix} x(\lambda_1) \\ x(\lambda_2) \end{bmatrix} &= \begin{bmatrix} k_A(\lambda_1) & k_B(\lambda_1) \\ k_A(\lambda_2) & k_B(\lambda_2) \end{bmatrix} \begin{bmatrix} y_A \\ y_B \end{bmatrix} \\ \begin{bmatrix} y_A \\ y_B \end{bmatrix} &= \begin{bmatrix} k_A(\lambda_1) & k_B(\lambda_1) \\ k_A(\lambda_2) & k_B(\lambda_2) \end{bmatrix}^{-1} \begin{bmatrix} x(\lambda_1) \\ x(\lambda_2) \end{bmatrix} \\ y_A &= b_A(\lambda_1)(\lambda_1) + b_A(\lambda_2)(\lambda_2) \\ y_B &= b_B(\lambda_1)(\lambda_1) + b_B(\lambda_2)(\lambda_2) \end{aligned} \tag{3}$$

Em termos gerais, o modelo multivariado é dado por

$$y = x_0 b_0 + x_1 b_1 + \ldots + x_{J-1} b_{J-1} + \epsilon \tag{4}$$

ou em notação vetorial,

$$Y = X\beta + \epsilon \tag{5}$$

com $x = [x_0\ x_1 \ldots x_{J-1}]$ é o vetor de valores medidos, $\beta = [b_0\ b_1 \ldots b_{J-1}]^T$ é o vetor a ser determinado e ϵ faz parte de erro aleatório.

Nesse caso de i amostras estarem disponíveis com n comprimento de onda, podemos organizar em pares $(x_i, y_i) \in \mathbf{R}^J \times \mathbf{R}$ tal que

$$Y = \begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_i^a \end{bmatrix} X = \begin{bmatrix} x_1^1(\lambda_1) & \ldots & x_1^j(\lambda_n) \\ x_2^1(\lambda_1) & \ldots & x_2^j(\lambda_n) \\ \vdots & \ddots & \vdots \\ x_i^1(\lambda_1) & \ldots & x_i^j(\lambda_n) \end{bmatrix}, \tag{6}$$

onde $x_i^j(\lambda_n)$ é a i-ésima amostra no comprimento de onda λ_n e y_i^a é a concentração de a na i-ésima amostra. Em que a relação entre a absorção e a concentração pode ser estimada por uma matriz de coeficientes β que multiplica X para se obter $\hat{Y}$ estimado. As matrizes X e Y são separadas em X_{cal} e Y_{cal} para obter a matriz de coeficientes β e X_{teste} e Y_{teste} são usados para testar a precisão do modelo de predição. Os coeficientes em β podem ser obtidos pelo modelo de regressão linear de acordo com a equação(7).

$$\beta = (X_{cal}^T X_{cal})^{-1} X_{cal}^T Y_{cal} \tag{7}$$

e $\hat{Y}$ pode ser estimado tal que

$$\hat{Y} = X_{teste}\beta \tag{8}$$

O problema acontece porque os dispositivos têm sido desenvolvidos para medir com mais precisão a absorbância, gerando um número cada vez maior de variáveis. Como consequência, há mais comprimentos de onda (variáveis) do que amostras (equações), no estudo de caso deste trabalho, por exemplo, temos 690 variáveis e 193 amostras na matriz X_{cal}.O s dispositivos mais modernos são capazes de gerar milhares de variáveis. Na Equação 7, se o número de variáveis é maior do que o número de amostras, a inversão não é possível ou mal condicionada.

Uma solução é a utilização de algoritmos de seleção de variáveis como algoritmos evolutivos para a escolha de um subconjunto de variáveis não redundante e sem colinearidade do conjunto original, ou a utilização de novas variáveis obtidas a partir de transformações lineares como o algoritmo de mínimos quadrados parciais (*Parcial Least Squares*, PLS).

Construir um modelo que inclui apenas um subconjunto de variáveis explicativas envolve dois objetivos conflitantes:

- Obter o máximo de informação por meio de um modelo com tantas variáveis independentes possíveis;
- Diminuir a variância da estimativa e o custo da coleta por meio de um modelo com menor número possível de variáveis.

E para tratar tais objetivos, o uso de algoritmos evolutivos é utilizado na resolução desse problema.

III. Algoritmos Evolutivos

Algoritmos Evolutivos (AE) é uma área de pesquisa da ciência da computação que se baseia no processo da evolução natural. Princípios como a mutação, recombinação genética, seleção natural e a sobrevivência do mais forte; são inspirações vindas da biologia que auxiliam na construção de algoritmos capazes de resolver problemas insolúveis para a computação. Um Algoritmo Evolutivo é um processo iterativo estocástico para gerar tentativas de solução para certo problema.

A aplicação dos Algoritmos Evolutivos em química foi inicialmente utilizada por Lucasius e Kateman [14] na seleção de comprimentos de onda em análises de sistemas multicomponentes. A partir disso, vários estudos tem utilizado essa abordagem na seleção de dados espectrofotométricos.

Conhecendo-se o problema proposto para utilização da regressão linear multivariada, é possível fazer com que um algoritmo evolutivo seja capaz de avaliar previamente os dados espectrofotométricos obtidos.

O algoritmo 1 apresenta um pseudo-código de um algoritmo evolutivo genérico. Esse algoritmo manipula um conjunto de indivíduos P (população), cada um dos quais compreende um ou mais cromossomos.

Algorithm 1 Pseudocódigo de um AE genérico.

$g = 0$;
$P(g) = Populacao\_Inicial(Pi(g_0))$;
$Avalia(Pi(g))$;
while $g \leq gmax$ **do**
  $Pi = Selecione\_Aleatoriamente(subpopulacao)$;
  $P' = Aplica\_Operadores(Pi(g))$;
  $Avalie(P')$;
  $P(g+1) = Sobreviventes(P(g), P')$;
  $g = g + 1$;
end while
return $min(P(gmax))$;

Os indivíduos mais aptos terão um maior número de descendentes, ao contrário dos indivíduos menos aptos. Os requisitos para um projeto de um AE são:

- Representações das possíveis soluções do problema no formato de um código genético;
- População inicial que contenha diversidade suficiente para permitir ao algoritmo combinar características e produzir novas soluções;
- Existência de um método para medir a qualidade de uma solução potencial;
- Um procedimento de combinação de soluções para gerar novos indivíduos na população;
- Um critério de escolha das soluções que permanecerão na população ou que serão retirados dessa;
- Um procedimento para introduzir periodicamente alterações em algumas soluções da população. Desse modo mantém-se a diversidade da população e a possibilidade de se produzir soluções inovadoras para serem avaliadas pelo critério de seleção dos mais aptos.

A. Algoritmo Evolutivo Multi-objetivo

Problemas de otimização multi-objetivo têm despertado grande interesse na área de Otimização. Nesses problemas, a qualidade da solução é definida com base na sua adequação em relação a diversos objetivos possivelmente conflitantes. Na prática, existem métodos de solução que buscam reduzir esses problemas a outros com apenas um objetivo para depois buscar uma solução global [15].

Uma classe de métodos bastante utilizada nesse contexto são os baseados em pesos ou simplesmente método de pesos, em que são atribuídos valores às diferentes funções objetivo, podendo ser realizado um processo de otimização da função de pesos.

Esses métodos, no entanto, são dependentes da escolha adequada dos pesos. Isso, em muitos casos, implica um conhecimento prévio dos intervalos correspondentes aos pesos mais adequados. Por essa razão, métodos que tentam encontrar soluções que apresentam um compromisso com os vários objetivos sem a utilização de pesos passaram a ser explorados [16]–[18].

IV. Algoritmo Evolutivo Multi-objetivo em Tabelas

O AE Multi-objetivo proposto trabalha com várias subpopulações em paralelo armazenadas em tabela, onde os melhores indivíduos para cada característica do problema são armazenados em sua respectiva subpopulação. Por essa razão, o AE desenvolvido é chamado de AE Multi-objetivo em Tabela em que seu pseudocódigo pode ser observado no algoritmo 2.

Cada subpopulação armazena indivíduos de modo a minimizar características como: (i) predição de erro (RMSEP), (ii) número de variáveis, (iii) sensibilidade à ruídos externos através da norma do vetor de coeficientes e (iv) função de agregação que pondere os outros objetivos.

O quarto objetivo é necessário para que seja considerado indivíduos que ponderem entre os objetivos considerados. Em razão das grandezas dos objetivos citados serem diferentes, é necessário a implementação de uma função agregadora, que utiliza de um mecanismo de auto-escalonamento para tratar as grandezas de modo uniforme.

Definido os parâmetros de execução para o algoritmo, gera-se uma população inicial P, com quantidade de indivíduos Qi. Os indivíduos em P são avaliados e separados em subpopulações de acordo com $vfit$, que é um vetor com os valores de aptidão de cada objetivo considerado. Feita essa separação inicial, o ciclo evolutivo se inicia, repetindo-se pelo número máximo de gerações $gmax$.

O método do torneio é utilizado para seleção, em que consiste na escolha aleatória de uma tabela e dessa, 3 indivíduos para sorteio, dentre esses o que tiver melhor aptidão é o escolhido. Esse processo é executado primeiro para o pai e em seguida para a mãe.

Dois indivíduos podem ser escolhidos de qualquer subpopulação da tabela, aumentando a diversidade entre os indivíduos que reproduzem de forma que as características dos indivíduos de uma subpopulação possam migrar para as demais

Algorithm 2 Pseudocódigo do AE Multi-objetivo em Tabelas.

```
P = 0;                                          ▷ População
SubPops = 0;                                    ▷ Tabela de Objetivos
vfit = 0;
Seja Qi a quantidade de indivíduos
P = gera_populacao_inicial(Qi);
for i = 1 to Qi do
    individuo = P(i);
    vfit = avalia(individuo);
    inserir_subpop(SubPops, vfit, individuo);
end for
for g = 0 to gmax do
    for tam = 1 to size(SubPops) do
        pai, mae = torneio(SubPops);
        if rand > pc then
            novos_filhos = cruzamento(pai, mae);
        end if
        if rand > pm then
            Seja Qfg, quantidade de filhos gerados
            for i = 1 to Qfg do
                mutacao(novos_filhos(i));
            end for
        end if
    end for
    for j = 1 to Qfg do
        individuo = novos_filhos(j);
        vfit = avalia(individuo);
        inserir_subpop(SubPops, vfit, individuo);
    end for
end for
```

subpopulações da tabela. Em consequência, aumenta-se a possibilidade do algoritmo escapar de ótimos locais, aproximando-se de soluções com avaliações próximas de um ótimo global. O cruzamento entre esses indivíduos e a mutação ocorrerá se a probabilidade de cruzamento pc ou se a probabilidade de mutação pm for atendida. A quantidade de filhos a serem gerados é dado por Qfg.

Após a avaliação, o indivíduo pode ser descartado ou armazenado em uma da tabelas pelo método $inserir\_subpop$, dependendo do grau de adaptação para cada objetivo do problema. O mesmo indivíduo pode ser armazenado em mais de uma tabela de acordo com esse critério.

V. Materiais e Métodos

Para a implementação do algoritmo genético foram utilizados o método de torneio para seleção, cruzamento uniforme e mutação para cada gene foram utilizados como modificadores genéticos.O número de cromossomos utilizado para a população inicial é de 120 indivíduos, com um número de 100 gerações, a probabilidade de cruzamento utilizada é de 70% e de mutação do indivíduo de 50%, com chance de mutação de 30% para cada gene. Foram feitas 50 execuções da implementação que resultaram em 50 arquivos para se obter a média dos resultados gerados. O uso de uma taxa de mutação mais alta que o usual é pelo fato de que, como um mesmo indivíduo pode estar presente em mais de uma tabela ao mesmo tempo, é necessário prevenir a convergência prematura dos indivíduos nas tabelas. Após alguns testes, foi possível perceber que aumentando a probabilidade de mutação em até 50%, foi possível diminuir consideravelmente essa convergência.

Para esse trabalho, o conjunto de dados utilizado para experimentação consiste em amostras de espetro fotométricas do trigo 775 VIS-NIR, que também foram utilizados como dados de exemplo na *International Diffuse Reflectance Conference* de 2008. Valores de proteína foram escolhidos como propriedade de interesse.

Os objetivos considerados são o RMSEP, o número de variáveis e a norma 2 do coeficientes de regressão. O erro quadrático médio obtido com o conjunto de predição (*Root Mean Square Error of Prediction*, RMSEP) é definido como:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{9}$$

em que $\hat{y}_i$ é a i-ésimo valor estimado pelo modelo, y_i é o i-ésimo valor real da concentração de interesse e N é a quantidade de amostras utilizadas no modelo.

O número de variáveis é dado por um escalar indicando a quantidade de variáveis utilizadas no modelo. A norma 2 do coeficientes de regressão ($||\mathbf{b}||$) é definida como:

$$||\mathbf{b}|| = \sqrt{\sum_{k=1}^{K}(\mathbf{b}(k))^2} \tag{10}$$

em que K é a quantidade de variáveis utilizadas no modelo. Maiores informações sobre o uso da norma 2 e sua relação com a sensibilidade à ruído podem ser encontradas em [9].

A configuração computacional deste trabalho consiste em um computador de mão equipado com processador Intel® Core i5, 6 GB de memória RAM. Matlab® 8.0.0.783 (R2012b) foi o software utilizado para implementação desse trabalho.

VI. Resultados e Discussões

Primeiramente, apresenta-se na Tabela I os resultados dos algoritmos clássicos PLS, SPA-MLR e MONO-GA-MLR. Como pode ser observado, os valores de RMSEP são similares para os três algoritmos nos resultados obtidos. No entanto, o MONO-GA-MLR seleciona um número excessivo de variáveis quando comparado ao SPA-MLR. Esse resultado pode ser explicado pelo fato de que o MONO-GA-MLR utiliza somente um objetivo, o RMSEP do conjunto de validação.

Assim, quando a inclusão de uma variável minimiza o RMSEP, por menor que seja, ela é incluída no modelo pelo MONO-GA-MLR. Na prática, o SPA-MLR é utilizado, pois utiliza menos variáveis que o MONO-GA-MLR e PLS. Vale a pena observar que o PLS utiliza todas as variáveis originais para construir as novas variáveis latentes. Os próximos parágrafos apresentam os resultados obtidos pelo algoritmo proposto, AEMT-MLR.

O AEMT-MLR, proposto neste trabalho, obteve os resultados apresentados na Tabela II. Ao compararmos os resultados do algoritmo mono-objetivo com o resultado do AEMT-MLR, é possível dizer que o algoritmo proposto obteve um resultado

Tabela I. RESULTADOS DAS TÉCNICAS TRADICIONAIS PLS, SPA-MLR AND MONO-GA-MLR. OS RESULTADOS ESTÃO EXPRESSOS EM VALORES DE RMSEP.

	RMSEP	Número de Variáveis
PLS	0.21	15*
SPA-MLR	0.20	13
MONO-GA-MLR	0.21	146

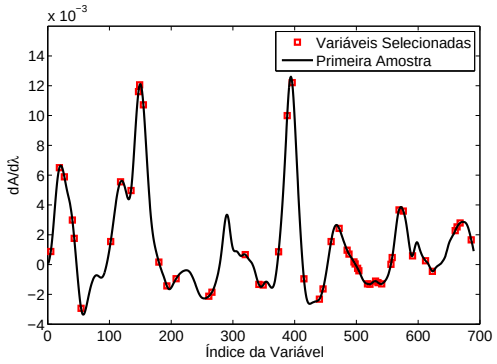
*Número de variáveis latentes.

Tabela II. RESULTADO DO ALGORITMO PROPOSTO.

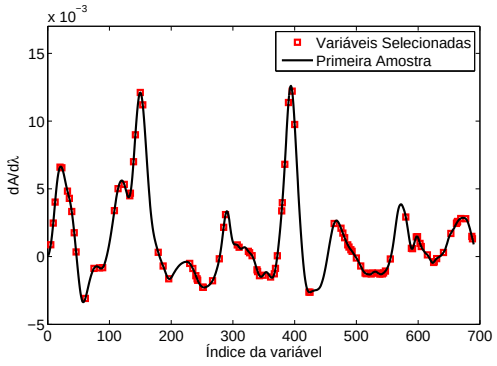
	RMSEP	Número de Variáveis
AEMT-MLR	0.0562	50

melhor que o mono-objetivo, tanto em relação ao número de variáveis quanto ao valor de RMSEP.

Observa-se na Figura 2 que o AEMT-MLR seleciona uma quantidade de variáveis significantemente menor que a quantidade do MONO-GA-MLR, no entanto os algoritmos possuem variáveis em regiões semelhantes do espectro. Isso indica que provavelmente essas são as regiões de interesse do espectro para o analito analisado.



(a)



(b)

Figura 2. Espectro da Amostra e Variáveis Selecionadas pelo (a) AEMT-MLR e (b) MONO-GA-MLR.

Na Figura 3 compara-se a sensibilidade à ruídos entre o algoritmo proposto e os algoritmos clássicos. Pode-se observar que a solução proposta pelo AEMT-MLR atingiu valores melhores de robustez se comparado ao algoritmo MONO-GA-MLR. O uso de apenas um objetivo no MONO-GA-MLR conduz o algoritmo à uma situação de *overfitting* e consequentemente à uma capacidade de generalização menor do que a formulação multi-objetivo.

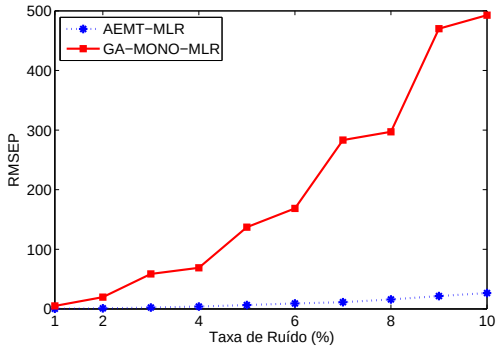


Figura 3. Análise da Taxa de Ruído entre AEMT-MLR e MONO-GA-MLR.

Na Figura 4 é apresentado a sensibilidade dos melhores indivíduos de cada tabela do AEMT-MLR, sendo: Tabela 1, minimização do RMSEP; Tabela 2, minimização do número de variáveis; Tabela 3, minimização da Norma 2; Tabela 4, função de ponderação dos três anteriores. Como pode ser observado o cromossomo obtido a partir da Tabela 1 apresenta o pior desempenho na presença de ruído.

Entretanto, para os demais cromossomos possuem um melhor desempenho quando comparado ao algoritmo de referência (SPA-MLR), com destaque para o cromossomo da Tabela 3 que apresentou a menor sensibilidade entre os todos os casos considerados. Tal resultado corrobora a premissa de que a norma 2 minimiza os efeitos indesejados do modelo em relação à presença de ruído instrumental.

Embora o melhor desempenho à sensibilidade ao ruído esteja nos cromossomos da Tabela 3 é possível inferir que as informações genéticas que fazem com que os cromossomos de tal tabela possuam uma menor sensibilidade foram propagadas para as demais tabelas de forma significativa, exceto para os cromossomos da Tabela 1.

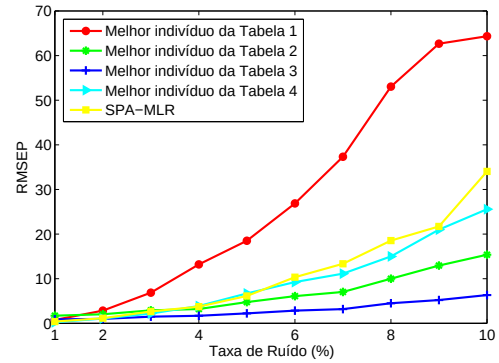


Figura 4. Análise da Taxa de Ruído entre os melhores indivíduos de cada tabela do AEMT-MLR e SPA-MLR.

A Figura 5(a) mostra as soluções da população final no espaço de busca dos três objetivos considerados. Nas Figuras 5(b), 5(c) e 5(d) apresenta-se as soluções considerando os objetivos dois a dois. Observa-se que quando são consideradas

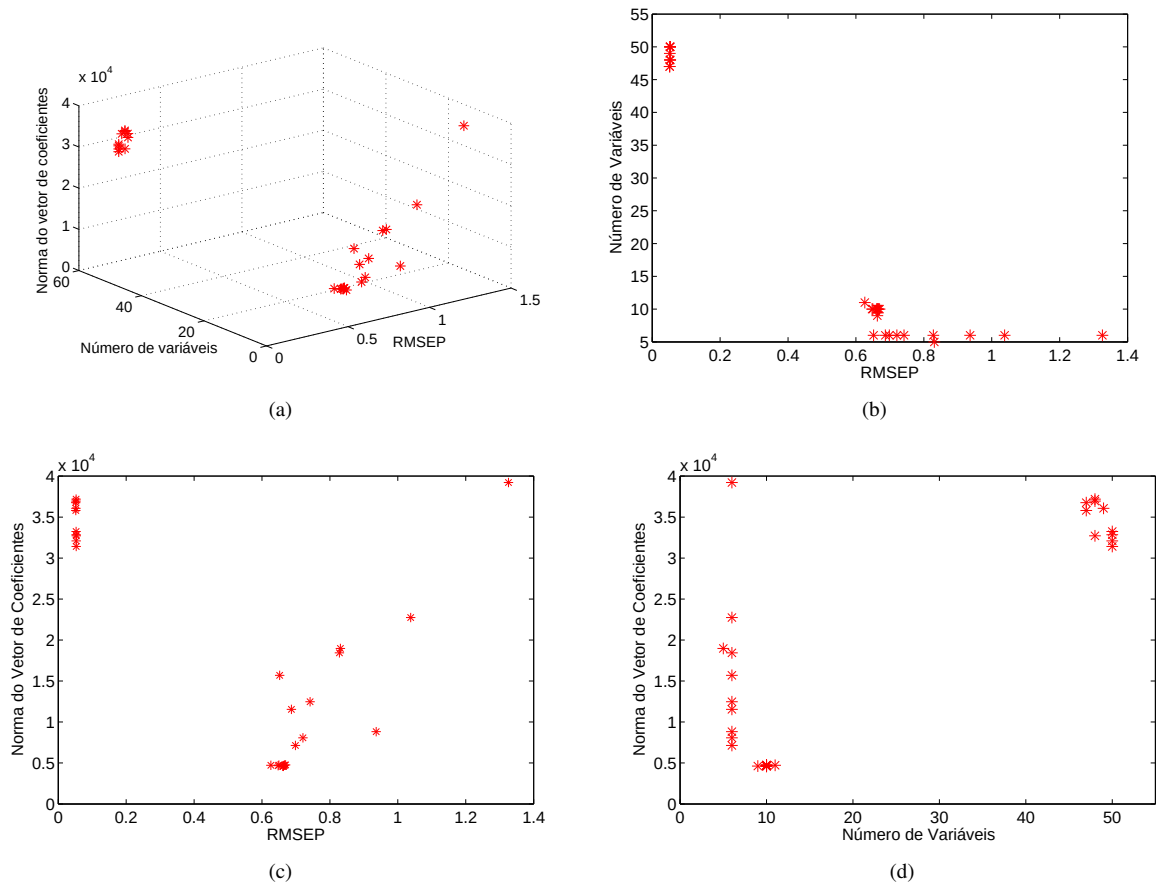


Figura 5. (a) Fronteira de Pareto do Algoritmo Proposto, (b) Relação RMSEP por Número de Variáveis, (c) Relação RMSEP por Norma do Vetor de Coeficientes, (d) Relação Número de Variáveis por Norma do Vetor de Coeficientes.

os objetivos de número de variáveis e RMSEP, as soluções da população final formam uma fronteira distinta. No entanto, quando o objetivo da norma 2 é utilizado comparando os objetivos dois a dois, as soluções não formaram uma fronteira distinta. Dessa forma, é necessária a utilização dos três objetivos para obter uma população totalmente não-dominada, uma vez que a norma 2 não consegue essa distinção.

VII. Conclusão

Este trabalho propôs uma nova implementação de algoritmo evolutivo para seleção de variáveis em problemas de calibração multivariada. Em particular, foi proposto um novo algoritmo denominado Algoritmo Evolutivo Multiobjetivo em Tabelas (AEMT) que visa facilitar o uso de três ou mais objetivos. Para o problema em estudo foi proposto o uso de quatro tabelas em que as três primeiras contém os seguintes objetivos: 1) minimização do erro de predição RMSEP, 2) número de variáveis utilizadas no modelo de regressão e 3) minimização da norma 2 do vetor de coeficientes de regressão. O quarto objetivo preserva indivíduos que ponderem os três objetivos anteriormente citados. Como exemplo de aplicação, um problema de calibração multivariada em dados de espectometria de trigo para determinação da quantidade de proteína foi utilizado.

Os resultados demonstraram que foi possível obter um erro aproximado substancialmente menor quando comparado ao algoritmo genético clássico mono-objetivo e a outros algoritmos clássicos de seleção de variáveis. Adicionalmente demonstrou-se que o uso da norma 2 como um dos objetivos considerados, leva o modelo de regressão a uma menor sensibilidade à presença de ruído instrumental.

Agradecimento

Os autores agradecem à Universidade Federal de Goiás, pela oportunidade de desenvolver esse trabalho. Agradecem também à CAPES e à FAPEG (Fundação de Amparo à Pesquisa do Estado de Goiás) pelo fomento provido para o desenvolvimento deste trabalho.

VIII. Apêndice - Relação entre Norma 2 e Sensibilidade à ruído

No método de mínimos quadrados não se consideram incertezas sobre os valores das variáveis independentes. Contudo em problemas de calibração multivariada para análises espectrofotométricas, deve-se ter em mente que os valores das variáveis independentes correspondem a medidas instrumentais contaminadas por ruído.

Desse modo convém ter uma medida da sensibilidade das predições do modelo com respeito a tal ruído instrumental

[9]. Neste apêndice, mostra-se que tal medida de sensibilidade pode ser dada pela norma-2 do vetor de coeficientes da regressão. De fato, considere que um modelo da forma

$$\hat{y} = b_0 + \sum_{k=1}^{K} b_k x_k \tag{11}$$

seja utilizado para calcular valores $\hat{y}$ previstos a partir de valores medidos $x = [x_1, x_2, ..., x_k]^T$. Além disso, suponha que tais valores medidos estejam contaminados por ruído, de modo que

$$x = \mu + \eta \tag{12}$$

sendo $\mu = E\{x\}, E\{\eta\} = 0$ e $E\{\eta\eta^T\} = \Sigma_\eta$, onde $E\{\cdot\}$ denota o valor esperado de uma variável aleatória. Tem-se portanto

$$\begin{aligned} E\{\hat{y}\} &= b_0 + E\left\{\sum_{k=1}^{k} b_k x_k\right\} \\ &= b_0 + \sum_{k=1}^{k} b_k E x_k \\ &= b_0 + \sum_{k=1}^{k} b_k \mu_k 0 + \sum_{k=1}^{k} b_k \mu_k \end{aligned} \tag{13}$$

Assim, a variância de $\hat{y}$ pode ser obtida por

$$\begin{aligned} \sigma^2_{\hat{y}} &= E\left\{[\hat{y} - E(\hat{y})]^2\right\} \\ &= E\left\{\left[\sum_{k=1}^{k} b_k(x_k - \mu_k)\right]^2\right\} \\ &= E\left\{\left[\sum_{k=1}^{k} b_k \eta_k\right]^2\right\} \\ &= E\{(b^T\eta)^2\} = E\{b^T\eta\eta^T b\} \\ &= b^T E\{\eta\eta^T\} b = b^T \Sigma_\eta b \end{aligned} \tag{14}$$

Se o ruído η for branco e homoscedástico, a matriz de variância-covariância Σ_η é da forma $\sigma^2_\eta I$, onde I e uma matriz identidade (KxK). Neste caso, a Equação (14) torna-se

$$\sigma^2_{\hat{y}} = \sigma^2_{\hat{y}} b^T b = \sigma^2_{\hat{y}} \|b\|_2^2 \tag{15}$$

onde $\|b\|_2$ é a norma-2 do vetor de coeficientes de regressão. A Equação (15) mostra que vetores de regressão maiores (no sentido da norma-2) tendem a gerar previsões que são mais sensíveis ao ruído de medição nas variáveis x.

REFERÊNCIAS

[1] M. M. C. Ferreira, A. M. Antunes, M. S. Melgo, and P. L. O. Volpe, "Quimiometria i: calibração multivariada, um tutorial," *Química Nova*, vol. 22, pp. 724 – 731, 09 1999.

[2] E. A. Coelho-Barros, P. A. Simões, J. A. Achcar, E. Z. Martinez, and A. C. Shimano, "Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos," *Revista Colombiana de Estadística, Vol. 31, 2008*, pp. no. 1, pp. 111 a 129, 2008.

[3] M. Arakawa, Y. Yamashita, and K. Funatsu, "Genetic algorithm-based wavelength selection method for spectral calibration," *Journal of Chemometrics*, vol. 25, no. 1, pp. 10–19, 2011.

[4] E. D. Moreira, M. J. Pontes, R. K. Galvão, and M. C. Araújo, "Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection," *Talanta*, vol. 79, no. 5, pp. 1260–1264, 2009.

[5] A. M. S. T. G. R. Y. T. C. H. V. V, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.

[6] S. F. Soares, A. A. Gomes, M. C. Araujo, A. R. G. Filho, and R. K. H. Galvao, "The successive projections algorithm," *TrAC Trends in Analytical Chemistry*, vol. 42, p. 84–98, 2013.

[7] D. V. de Lucena, T. W. de Lima, A. da Silva Soares, A. R. G. Filho, and C. J. Coelho, "Multiobjective evolutionary algorithm for variables selection in calibration problems: a case study for protein concentration prediction," *Proccedings of the IEEE Congress of Evolutionary Computation*, 2013.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast elitist multi-objective genetic algorithm: Nsga-ii," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 182–197, 2000.

[9] A. R. Galvão Filho, M. Araújo, and R. K. H. Galvão, "Effect of the subsampling ratio in the application of subagging for multivariate calibration with the successive projections algorithm," *Journal of the Brazilian Chemical Society*, vol. 22, pp. 2225–2233, 2011.

[10] D. A. Skoog, *Princípios de análise instrumental*. Bookman, 2002.

[11] M. Shimada, Y. Masuda, Y. Yamada, M. Itoh, M. Takahashi, and T. Yatagai, "Explanation of human skin color by multiple linear regression analysis based on the modified lambert-beer law," *Optical Review*, vol. 7, no. 4, pp. 348–352, 2000.

[12] R. H. Guy, J. J. Hostynek, R. S. Hinz, and C. R. Lorence, *Metals and the Skin*. Marcel Dekker Incorporated, 1999.

[13] T. Naes and B. H. Mevik, "Understanding the collinearity problem in regression and discriminant analysis," *Journal of Chemometrics*, vol. 15, no. 4, pp. 413–426, 2001.

[14] C. B. Lucasius and G. Kateman, "Genetic algorithms for large-scale optimization in chemometrics - an application, trends in analytical chemistry," *TrAC Trends in Analytical Chemistry*, vol. 10, pp. p.254–261, 1991, 1991.

[15] D. T. da Silva, "Algoritmos genéticos e o problema de corte multiobjetivo," Master's thesis, Universidade Estadual de Campinas, São Paulo, may 2009.

[16] A. M. Lopes, "Uma abordagem multiobjetivo para o problema de corte de estoque unidimensional," Master's thesis, Universidade Estadual Paulista Júlio de Mesquita Filho, Jan. 2009.

[17] R. O. Parreiras, "Algoritmos evolucionários e técnicas de tomada de decisão em análise multicritério," 2006.

[18] D. S. Sanches, "Algoritmos evolutivos multi-objetivo para reconfiguração de redes em sistemas de distribuição de energia elétrica," Universidade de São Paulo, São Paulo, dec 2012.