PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PROGRAMA DE MESTRADO EM ENGENHARIA DE PRODUÇÃO
E SISTEMAS

ALGORITMO CO-EVOLUTIVO PARA PARTICIONAMENTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMA DE CALIBRAÇÃO MULTIVARIADA

JORCIVAN SILVA RAMOS

2015

ALGORITMO CO-EVOLUTIVO PARA PARTICIONAMENTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMA DE CALIBRAÇÃO MULTIVARIADA

JORCIVAN SILVA RAMOS

Dissertação de Mestrado apresentada ao Programa de Pós–Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Mestre em Programa de Mestrado em Engenharia de Produção e Sistemas.

Orientador: Prof. Dr. Clarimar José Coelho
Co-Orientador: Prof. Dr. Anderson da Silva Soares

GOIÂNIA
AGOSTO DE 2015

Dados Internacionais de Catalogação da Publicação (CIP)
(Sistema de Bibliotecas PUC Goiás)

Ramos, Jorcivan Silva.
R175a Algoritmo co-evolutivo para particionamento de dados e seleção de variáveis em problema de calibração multivariada [manuscrito] / Jorcivan Silva Ramos – Goiânia, 2015.
11 f. : il. ; 30 cm.

Dissertação (mestrado) – Pontifícia Universidade Católica de Goiás, Programa de Pós-Graduação *Stricto Sensu* em Engenharia de Produção
e Sistemas, 2015.
"Orientador: Prof. Dr. Clarimar José Coelho".
Bibliografia: p. 43-50.

1. Algorítmos genéticos. I. Título.

CDU 004.421(043)

ALGORITMO CO-EVOLUTIVO PARA PARTICIONAMENTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMA DE CALIBRAÇÃO MULTIVARIADA

Jorcivan Silva Ramos

Esta Dissertação julgada adequada para obtenção do título de Mestre em Engenharia de Produção e Sistemas e aprovada em sua forma final pelo Programa de Pós–Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás em AGOSTO de 2015.

_____

Prof. Dr. Ricardo Luiz Machado
Coordenador do Programa de Pós-Graduação em Engenharia de Produção e Sistemas

Banca Examinadora:

_____

Prof. Dr. Clarimar José Coelho
Engenharia de Produção e Sistemas – PUC-GO
Orientador

_____

Prof. Dr. Alexandre Cláudio Botazzo Delbem
Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo – USP

_____

Prof. Dr. Ricardo Luiz Machado
Engenharia de Produção e Sistemas– PUC-GO

GOIÂNIA-GOIÁS
AGOSTO DE 2015

Agradecimentos

- A Deus pelo fôlego de vida.
- A toda minha família, mãe, pai, esposa e irmãos pelo apoio.
- Ao prof. Dr. Anderson Soares e sua esposa Telma Soares, pelas orientações e por sempre estarem disponíveis.
- Ao prof. Dr. Clarimar, pelas suas orientações e ensinamentos.
- Ao grupo do Laboratório de Análise Multivariada (LAMV) da PUC-GO, pelas várias oportunidades de crescimento profissional.
- Ao Instituto Federal Goiano pelo apoio institucional.

Resumo da Dissertação apresentada ao MEPROS/ PUC Goiás como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia de Produção e Sistemas (M.Sc.)

ALGORITMO CO-EVOLUTIVO PARA PARTICIONAMENTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMA DE CALIBRAÇÃO MULTIVARIADA

JORCIVAN SILVA RAMOS

Agosto/2015

Orientador: Dr. Clarimar José Coelho

Esse trabalho apresenta o desenvolvimento de um algoritmo genético co-evolutivo para a seleção de amostras a partir de um conjunto de dados e a seleção de variáveis a partir das amostras selecionadas no contexto da calibração multivariada. Cada amostra é dividida em conjunto de calibração para a confecção do modelo e conjunto de validação do modelo de calibração. O algoritmo seleciona amostras e variáveis com o objetivo de construir modelos de calibração. Os resultados mostram que os conjuntos de dados selecionados pelo algoritmo proposto produzem modelos com melhor capacidade preditiva do que os modelos relatados na literatura.

Palavras–chave
Algoritmo genético; co-evolução; calibração multivariada; seleção de amostras; seleção de variáveis.

ABSTRACT

ALGORITMO CO-EVOLUTIVO PARA PARTICIONAMENTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMA DE CALIBRAÇÃO MULTIVARIADA

JORCIVAN SILVA RAMOS

Agosto/2015.

Orientador: Dr. Clarimar José Coelho

This paper presents the development of a co-evolutionary genetic algorithm for the selection of samples from a data set and the selection of variables from the samples selected in the context of multivariate calibration . Each sample is divided into the calibration set for the preparation of the model and validating the calibration set of model. The algorithm selects samples variables with the goal of building the calibration models. The results show that the data sets selected by the proposed algorithm models to produce better predictive ability of the models reported in the literature.

Keywords
Genetic Algorithm; co-evolution; multivariate calibration; selection of samples; selection of variables.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

d**x** (*p*,*q*) - Distância euclidiana do Algoritmo Kennard-Stone

d**y** (*p*,*q*) - Distância euclidiana do Algoritmo SPXY

j - Índice do vetor

K - Quantidade de amostras

Kv - Número de amostras do conjunto de validação

x - Vetor das variáveis independentes

X - Matrix de variáveis e amostras

y - Vetor das variáveis dependentes

y_v^K - Valor de referência

$\hat{y}_v^K$ - Valor de previsão do parâmetro de interesse

Lista de Abreviaturas e Siglas

AGCP - Algoritmo Genético Co-evolutivo Proposto

GA - Algoritmo genético (*Genetic Algorithm*)

GHz - *Gigahertz*

RNG - Gerador de números aleatórios (*Random Number Generator*)

K-S - Kennard-Stone

NIR - Infravermelho próximo (*Near-Infrared*)

PLS - Mínimos quadrados parciais (*Partial Least Squares*)

PRESS - Soma dos quadrados dos erros de previsão (*Prediction Error Sum of Squares*)

RAM - Memória de acesso aleatório (*Random Access Memory*)

RMSEP - Erro médio quadrático de predição (*Root Mean Square Error of Prediction*)

SGA - Algoritmo genético simples (*Simple Genetic Algorithm*)

SPA - Algoritmo de projeções sucessivas (*Sucessive Projections Algorithm*)

SPXy - Partição do conjunto de amostras baseado nas distâncias conjuntas (*Sample set Partitioning based on join X-y distances*)

1. Introdução

Algoritmos inspirados na evolução biológica para a solução de problemas computacionais são chamados algoritmos evolutivos (BREMERMANN *et al.* 1966, FRASER, 1960). Algoritmos genético co-evolutivos são tipos de algoritmos evolutivos em que a co-evolução é associada à evolução mútua entre duas instâncias do algoritmo que apresentam dependência entre si e uma instância exerce pressão seletiva sobre a outra (PENA-REYES e SIPPER, 2001). Tipicamente, algoritmos evolutivos, são usados para a solução de problema de otimização combinatória com uma função objetivo com crescimento ou desenvolvimento populacional, busca aleatória guiada e processamento paralelo (SUMATHI, *et al.*, 2008).

Este trabalho apresenta o desenvolvimento e os resultados obtidos com a aplicação de um algoritmo co-evolutivo para seleção de amostras e a seleção de variáveis a partir de um conjunto de dados de trigo proveniente de métodos instrumentais (SKOOG, *et al*., 1997). Tem como objetivo encontrar o melhor conjunto de amostras e o melhor conjunto de variáveis para a confecção de modelos de validação cruzada e calibração multivariada (MARTENS, 2001).

A calibração multivariada consiste no algoritmo de calibração ou modelo de calibração usado para obter as propriedades de interesse a partir de valores registrados por métodos instrumentais (WOLD,1995). Quando a relação é estabelecida entre mais de duas variáveis o processo é denominado de calibração multivariada (NAES *et al*., 2002).

O processo de calibração multivariada requer o uso de dois conjuntos de dados denominados conjunto de calibração e conjunto de predição. A divisão de subconjuntos de amostras deve ser feita de modo a garantir que as amostras mais representativas estejam no conjunto de calibração (BROWNE e CUDECK, 1989). O conjunto de calibração é utilizado para guiar o algoritmo a obter as propriedades de interesse entre a resposta instrumental (variáveis independentes) e a propriedade de

interesse (variáveis dependentes) (MARTENS, 2001) e conjunto de predição que é usado para prever a propriedade de interesse. A divisão de subconjuntos de amostras deve ser feita de modo a garantir que as amostras mais representativas estejam no conjunto de calibração (BROWNE e CUDECK, 1989). O conjunto de validação ou teste é utilizado para avaliar a qualidade do modelo obtido a partir do conjunto de calibração. O algoritmo de calibração é usado para prever a propriedade de interesse e comparar o valor predito com valores certificados.

Uma alternativa ao processo de divisão e formação dos conjuntos de calibração e validação é a utilização da validação cruzada que em geral os dados são provenientes de métodos instrumentais que contam com evoluídos instrumentos computadorizados que permitem a coleta de grandes quantidades de amostras (C. ANTUNES A. M., 1999). A validação cruzada efetua testes completos sobre a variação do modelo em relação aos dados utilizados. Quando se trata de um conjunto grande de amostras é necessário dividir os dados em conjuntos menores de amostras no processo de criação do algoritmo de calibração (LI *et al*., 2009). A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados (WOLD, 1995).

Uma das dificuldades da construção de modelos de calibração multivariada é a escolha das amostras que representam a maior variância das características em estudo. A inclusão de maior variabilidade facilita a construção de modelos com maior capacidade preditiva (NAES e MEVIK, 2001).

O algoritmo Kennard-Stone (K-S) foi desenvolvido em 1969 e é usado para selecionar as amostras de calibração em um conjunto de dados. O K-S utiliza como critério de seleção a maior distância entre as amostras já selecionadas, tentando assegurando uma distribuição uniforme na seleção das amostras. (KENNARD, 1969).

O K-S é utilizado por Sousa (2008) para selecionar amostras para a predição de características da madeira para a produção de celulose. Neves (2013) usa o K-S para a confecção de algoritmos de calibração multivariada aplicados à determinação

simultânea de parâmetros bioquímicos em plasma sanguíneo. Santos *et al*., (2015) utiliza o K-S para a seleção de amostras em problemas de calibração multivariada para a determinação da acidez total em bebidas industrializadas a base de soja e néctar de frutas.

O algoritmo de partição do conjunto de amostras baseado nas distâncias conjuntas **X-y** (*Sample set Partitioning based on join* **X-y** *distances*, SPXy) (GALVÃO *et al.,* 2005) é uma extensão do algoritmo K-S que atribui aumento da distância entre as amostras e leva em consideração tanto as diferenças das variáveis independentes ou regressoras **X** quanto as diferenças da variável **y** (propriedade de interesse) no cálculo das distâncias. O SPXy considera de igual importância as distribuições das amostras no espaço **X** e **y**.

A seleção de variáveis em amostras para a escolha da região espectral que contém informações mais representativas e com menor redundância para a propriedade de interesse pode ser uma estratégia importante para a melhoria da qualidade dos algoritmos de calibração (LEARDI, 2000), sendo necessária, uma vez que as técnicas de calibração multivariada ficam limitadas quando existe um número relativamente elevado de variáveis no conjunto de dados (VASCONCELOS, 2011). A seleção de variáveis contribui para a construção de modelos de calibração simples e robustos, além de minimizar erros de predição (PASQUINI 2003, OLIVEIRA *et al*. 2004).

Uma técnica que tem sido muito utilizado para seleção de variáveis, principalmente em dados espectrais de análises químicas é o algoritmo de quadrados mínimos parciais (*Partial Least Squares,* PLS), cuja ideia central é a extração das variáveis latentes para a construção do modelo. O algoritmo PLS estabelece uma decomposição simultânea da resposta instrumental (variáveis independentes) com o vetor de propriedade de interesse (variáveis dependentes) a partir dos componentes que representam o máximo da covariância entre eles (CHAUCHARD *et al*., 2004).

O PLS foi utilizado por Michel (ANZANELLO, 2009) visando à seleção de variáveis para fins de categorização de bateladas de produção em duas classes, a metodologia foi aplicada a três bancos de dados, caracterizados por relativamente altos de correlações entre as variáveis, o estudo apresentou resultados com redução grande no número de variáveis em relação a outras técnicas ferramentas testadas.

O algoritmo de projeções sucessivas (*Sucessive Projections Algorithm*, SPA) é bastante usado para a seleção de variáveis. Inicialmente seleciona um subconjunto de variáveis com mínima colinearidade, ou seja, as variáveis que possuem pequenas relações lineares exatas ou aproximadamente exatas e realiza operações de projeções aplicadas no conjunto de dados de calibração para seleciona um subconjunto de variáveis com base na capacidade de previsão. O SPA pode empregar a validação cruzada a partir da indicação do usuário. O processo é encerrado verificando se a variável pode ser excluída sem perdas de capacidade de previsão (ARAÚJO *et al.,* 2001).

O grande desafio e limitação desta técnica esta na dificuldade de interpretação, uma vez que houve uma transformação nas variáveis de entrada para a construção das variáveis latentes e que a regressão é realizada neste novo domínio dos dados, não há possibilidades de uma interpretação direta nos dados. É importante salientar que durante o processo de seleção das amostras e seleção de variáveis usando técnicas tradicionais em calibração multivariada, não são considerados os efeitos simultâneos de seleção de amostras e da seleção de variáveis. Este trabalho apresenta uma alternativa a esse problema que executa de forma concorrente a seleção de amostras e a seleção de variáveis de modo que uma exerça influência sobre a outra pelo uso do conceito da co-evolução.

O algoritmo genético co-evolutivo possui duas vantagens importantes sobre os algoritmos genéticos simples. A primeira é que os algoritmos genéticos simples não preservam algumas características de certos integrantes da solução, pois consideram o problema por completo, preservando apenas os indivíduos que tenham altas

avaliações. A segunda é que os algoritmos tradicionais estão relacionados a uma solução completa, não havendo interações entre os membros das populações, com isso, não existe a ocorrência de co-evolução. Desta forma, não há pressão para coadaptação de um subconjunto devido à mudança ocorrida no outro (POTTER e JONG, 2000).

Augusto (2009) propôs um algoritmo genético co-evolutivo para classificação de dados. A técnica foi explorada no nível inter-populacional, em que as populações cooperaram em um regime semi-isolado, e também no nível intra-populacional, em que os classificadores candidatos co-evoluíram competitivamente com as amostras de treinamento. O algoritmo teve melhor desempenho do que algoritmos tradicionais que não utilizam a co-evolução. Outra abordagem foi desenvolvida por Zuben (2002) que propõe um algoritmo genético co-evolutivo hierárquico para projetar sistemas baseados em regras nebulosas, em que os indivíduos cooperam representando diferentes parâmetros para o sistema. A abordagem nebulosa trata se de sistemas que se baseiam em codificação da informação (YAGER e FILEV, 1994).

Como estudo de caso será utilizado um conjunto de dados de trigo obtidos através da espectroscopia no infravermelho próximo (NIR) por reflectância difusa na faixa de 1100 a 2500 nm, com intuito de estimar a concentração de proteínas existentes. O teor de proteína no trigo é importante principalmente por conter um considerável fator nutricional. A quantidade e o tipo de proteína que são encontradas no trigo são de muita importância em sua utilização (ELIASSON e LARSSON, 1993).

O Capítulo 2 apresenta o principal problema a ser resolvido nesse trabalho e os dados usados para o teste do algoritmo proposto. A breve revisão dos métodos usados para a solução do problema introduz os algoritmos genéticos e seus elementos fundamentais como: função objetivo, indivíduo, seleção e reprodução. São introduzidos os fundamentos sobre algoritmos genéticos essenciais para o desenvolvimento do algoritmo genético coevolutivo proposto (AGCP) para a seleção de amostras e variáveis. O Capítulo 3 apresenta de maneira breve algumas técnicas

tradicionais usadas para a seleção de amostras em grandes conjuntos de dados e a seleção de variáveis tal como algoritmo K-S, algoritmo SPXy e SPA, PLS, respectivamente. O Capítulo 4 apresenta e analisa os resultados obtidos empregando o Algoritmo Genético Co-evolutvo Proposto (AGCP) frente a resultados presentes na literatura. O Capítulo 5 apresenta, analisa e crítica os resultados obtidos com o AGCP incluindo vantagens e limitações.

2. Materiais e Algoritmos Genéticos

Esse estudo está relacionado ao problema de selecionar a melhor amostra e as melhores variáveis a partir de um conjunto de dados para a confecção de algoritmos de calibração/modelos de calibração empregando algoritmos co-evolutivos de modo que o algoritmo de calibração tenha a melhor capacidade preditiva possível.

O conjunto de dados usados para o teste do algoritmo de calibração proposto é composto por espectros no infravermelho próximo na faixa de 400-2500 nm, com resolução de 2 nm, num total de 775 amostras extraídos de grãos de trigo inteiros obtidos no laboratório de pesquisa de grãos da cidade de Winnipeg no Canadá. Esse conjunto de dados foi utilizado como benchmark para os trabalhos apresentados na conferencia internacional em 2008 e estão disponíveis em (http://www.idrcchambersburg.org/shootout.html).

O laboratório fez o pré-processamento dos espectros da concentração de proteína com a aplicação da primeira derivada usando um filtro Savitzky-Golay com polinômio de 2ª ordem e uma janela de 11 pontos. O algoritmo Kennard-Stone (K-S) foi usado para dividir os dados em conjunto de validação, calibração e predição com um total de 690 variáveis independentes.

Em geral, as características de referência são a concentração de proteína (%), teste de peso (Kg), textura do grão de trigo (%), absorção de água por farinografia (%), tempo de desenvolvimento de massa por farinografia (em minutos). Para os objetivos desse trabalho, a concentração da proteína é a propriedade de interesse para o desenvolvimento do modelo de calibração (LASZTITY, 1995).

2.1 Algoritmo Genético

Algoritmo genético (*Genetic Algorithm*, GA) é uma técnica de busca de soluções aproximadas ou não determinísticas em problemas de otimização e busca inspirada na teoria da seleção natural de Darwin (HOLLAND, 1975). Na definição de GAs são usados termos da biologia evolutiva tais como hereditariedade, mutação, seleção natural e recombinação ou crossover (DARWIN, 1859).

Em GAs a população representa o conjunto de soluções de um problema e a evolução da população é iniciada a partir de um conjunto de soluções criadas aleatoriamente e feita por meio de gerações. A cada geração, a adaptação de cada solução da população é avaliada, alguns indivíduos são selecionados para a próxima geração e recombinados ou mutados para formar uma nova população. Um indivíduo é a representação do espaço de busca, todas as soluções possíveis de um problema ou a função matemática para o problema a ser resolvido. A população gerada é usada como entrada para a próxima iteração do algoritmo (GOLDBERG, 1989).

GAs são diferentes de outros algoritmos de otimização porque codificam um conjunto de soluções possíveis e não os parâmetros de otimização. O resultado é uma população de soluções e não uma solução única. O resultado produzido é avaliado por uma função matemática e não é necessário conhecimento prévio do problema. As transições são probabilísticas, pois dada uma mesma sequência de entrada, não necessariamente chega-se a um mesmo estado final (GASARCH, 1989).

Os componentes fundamentais de um GA são a função objetivo, o indivíduo, a seleção e a reprodução. O objeto da otimização em GAs é conhecido como função objetivo (função matemática do problema) e pode ser um problema de otimização, um conjunto de teste para obter indivíduos mais aptos ou problema do tipo caixa-preta onde são conhecidas as entradas e os resultados da otimização. Em GAs não é necessário saber como a função objetivo funciona, basta que ela esteja disponível

para ser aplicada aos indivíduos e os resultados comparados (LUCASIUS e KATEMAN, 1993).

Um indivíduo no contexto GA é o portador do código genético que representa o espaço de busca do problema a ser resolvido e pode ser representado por uma sequência de bits. Por exemplo, podem ser usados 8 bits para a otimização de problemas com valores inteiros positivos menores que 255. Problemas com múltiplas entradas podem combinar as entradas em uma única sequência de bits ou usar mais de um cromossomo onde cada cromossomo representa uma entrada. Cromossomo é a estrutura de dados que representa uma possível solução do problema como uma cadeia de bits ou vetores de inteiros. O código genético deve ser finito e capaz de representar todo o conjunto dos valores do espaço de busca (MITCHELL, 1998).

A seleção é o componente do GA que imita a seleção natural das espécies proposta por Darwin. O tipo mais comum de seleção de indivíduos é por seleção proporcional a aptidão e seleção por torneio. A roleta é um tipo de seleção proporcional que seleciona os indivíduos da população de acordo com a aptidão, os indivíduos ocupam uma porção da roleta proporcional a sua aptidão. Sendo assim, os que tiverem maior aptidão, terão maiores chances de serem sorteados pela roleta. A roleta é feita de forma aleatória e não garante que os melhores sejam escolhidos para a próxima geração (MICHALEWICZ e SCHOENAUER, 1996). Na seleção por torneio um subconjunto com k de indivíduos é construído. Em geral são sorteados dois indivíduos aleatoriamente e o que tiver uma melhor aptidão será escolhido. A vantagem do método é que não é necessário conhecimento global da população. A seleção elitista mantém para a próxima geração os melhores indivíduos. É a junção de muitos métodos que assegura que o GA mantenha alguns dos melhores indivíduos de cada geração (MITCHELL, 1998).

A reprodução é dividida em acasalamento, recombinação e mutação. O acasalamento consiste na escolha de dois indivíduos para gerar dois descendentes. A recombinação ou crossover imita o processo biológico onde os descendentes recebem

parte do código genético do pai e parte do código genético da mãe. O processo de recombinação garante que os melhores indivíduos troquem entre si as informações que os assegurem ser mais aptos para sobreviver e gerar descendentes ainda mais aptos. As mutações são feitas com a mais baixa probabilidade possível. Garante maior variabilidade genética na população e impede a estagnação da busca em um mínimo local (LUCASIUS e KATEMAN, 1993).

Métodos usuais de crossover são de um ponto de corte, dois pontos de corte e uniforme. A Figura 1 ilustra como é feito o método de um ponto de corte que consiste em dividir os dois cromossomos selecionados para serem os pais em um ponto que será escolhido aleatoriamente. Em seguida é copiada uma parte dos cromossomos de cada pai e gera dois filhos (LINDEN, 2008).

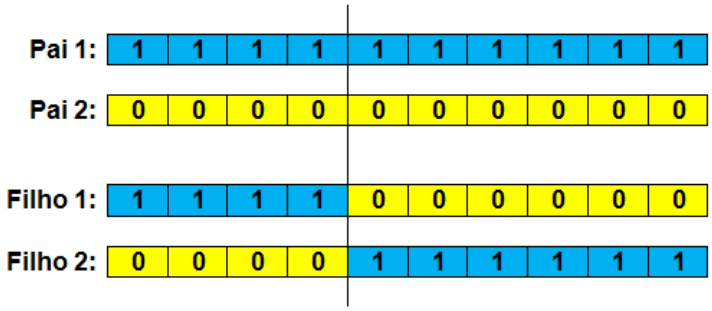

Figura 1: Cruzamento com um ponto de corte.

A Figura 2 ilustra o método de dois pontos de corte que é um processo similar ao método de um ponto de corte. Porém, aqui são feitos aleatoriamente dois cortes e o filho que receber a parte central dos cortes do cromossomo de um dos pais é formado na parte esquerda e na parte direita com as características do outro pai que fica intercalado (LINDEN, 2008).

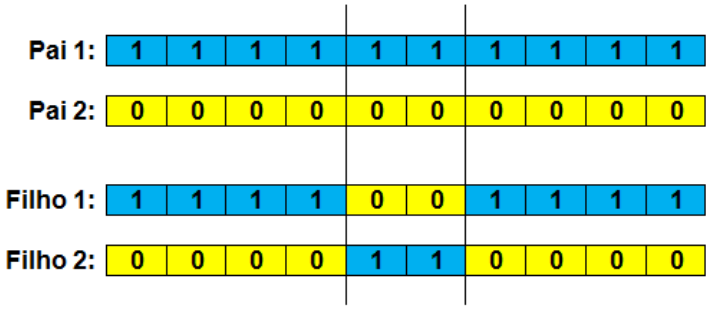

Figura 2: Cruzamento com dois pontos de cortes.

A Figura 3 mostra o funcionamento do método uniforme que não usa cortes. É criado aleatoriamente um cromossomo (máscara) que guia a troca de informações dos pais na geração dos filhos. Caso o valor do gene da máscara seja um, o primeiro filho recebe o gene do primeiro pai na posição corrente, já o segundo filho recebe o gene do segundo pai. Da mesma forma, se o valor da máscara é zero, o primeiro filho receberá o gene do segundo pai na posição corrente e o segundo filho recebe o gene do primeiro pai. Esse processo se repete até completar a quantidade dos genes dos filhos que estão sendo formados (LINDEN, 2008).

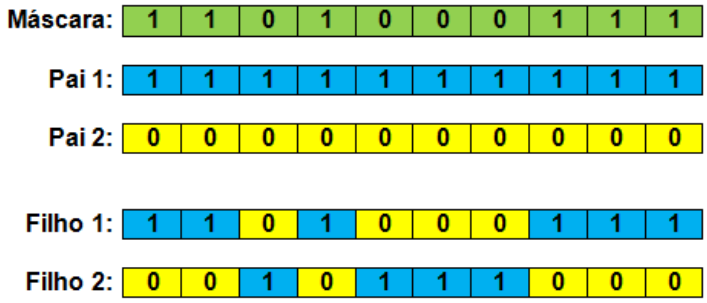

Figura 3: Cruzamento uniforme.

A mutação aleatoriamente modifica um ou mais genes dos indivíduos gerados. A mutação é responsável pela introdução e manutenção da diversidade genética da população (HOLLAND, 1975). Impede que o algoritmo sofra uma convergência prematura e também insere novos pontos no domínio de busca, por meio da inclusão

de características que não estão presentes na população. A taxa de mutação corresponde à probabilidade do operador de mutação ocorrer em um gene. Normalmente essa taxa terá valores pequenos, sendo que não há nenhuma garantia que a mudança tornará o cromossomo potencialmente melhor (GOLDBERG, 2002).

Os tipos de mutação mais usados em GAs são mutação flip, onde cada gene a ser alterado recebe um valor aleatório do alfabeto válido. Mutação *swap*, onde alguns pares de genes são sorteados e os pares entre si trocam os valores de seus genes. Mutação *creep* onde um valor aleatório é somado ou subtraído do valor do gene. Mutação *inversio*n, acontece quando a representação dos cromossomos é feita por codificação binaria, invertendo o valor do gene escolhido aleatoriamente de acordo com a taxa de mutação (LINDEN, 2008).

A probabilidade da mutação ocorrer em um gene deve ser estimada em torno de 0 a 10%, para que o processo de otimização não se torne puramente aleatório (SRINIVAS e PATNAIK, 1994). A Figura 4 mostra um exemplo de mutação sendo realizada com uma representação binária. A seta representa o gene que deve sofrer a mutação, onde seu valor deve ser invertido por se tratar de uma codificação binária, o gene selecionado tem o bit 1. Após a mutação passou a ter o bit zero (LINDEN, 2008).

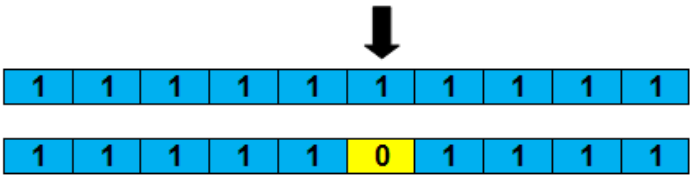

Figura 4: Mutação binária.

2.2 Algoritmos genéticos co-evolutivos

O termo co-evolução foi determinado por Ehrlich e Raven (1964) para descrever a influência que as plantas exercem na evolução de insetos herbívoros. Os autores mostram que existem associações entre algumas espécies de borboletas e as

plantas hospedeiras. Para Rosin e Belew (1997), a co-evolução é a evolução simultânea de duas ou mais espécies e a medida de desempenho da evolução é o resultado da evolução simultânea.

Um algoritmo co-evolutivo imita a extensão da evolução natural e busca solução de problemas a partir da interação de duas ou mais populações relacionadas que divide a solução de problemas complexos em subproblemas e contribui para a solução do problema como um todo. A abordagem co-evolutiva pode ser classificada como competitiva ou cooperativa (GREFENSTETTE e DALEY, 1996).

A evolução cooperativa resolve problemas complexos análogos a ecossistemas compostos por duas ou mais populações. Os cruzamentos dos indivíduos são realizados somente entre indivíduos da mesma população, evoluindo as espécies em populações separadas. Em um relacionamento cooperativo as espécies interagem entre si em um domínio compartilhado (POTTER e JONG, 2000). Cada indivíduo tem sua aptidão de acordo com a capacidade de cooperar com as outras espécies no desenvolvimento da solução global.

Um modelo co-evolucionário cooperativo básico é mostrado na Figura 5 onde as espécies evoluem em suas populações de acordo com o ambiente e através da repetição do algoritmo e formando espécies cada vez mais adaptadas ao meio. No modelo coevolutivo cooperativo mostrado na Figura 5 duas espécies evoluem separadamente até um determinado momento. A evolução é representada pelo bloco azul claro e indica que os indivíduos da espécie sofrem a evolução e são avaliados de acordo com a capacidade que cada um tem em solucionar um problema. No momento em que uma espécie para sua evolução a outra envia seus melhores colaboradores ao modelo de domínio e ocorre a junção de indivíduos de espécies diferentes em que uns cooperam com os outros na busca por uma melhor solução para o problema. A escolha por colaboradores pode ser feita de várias maneiras. A mais comum e também a mais simples é seleção do melhor individuo de cada população (IORIO e LI, 2004).

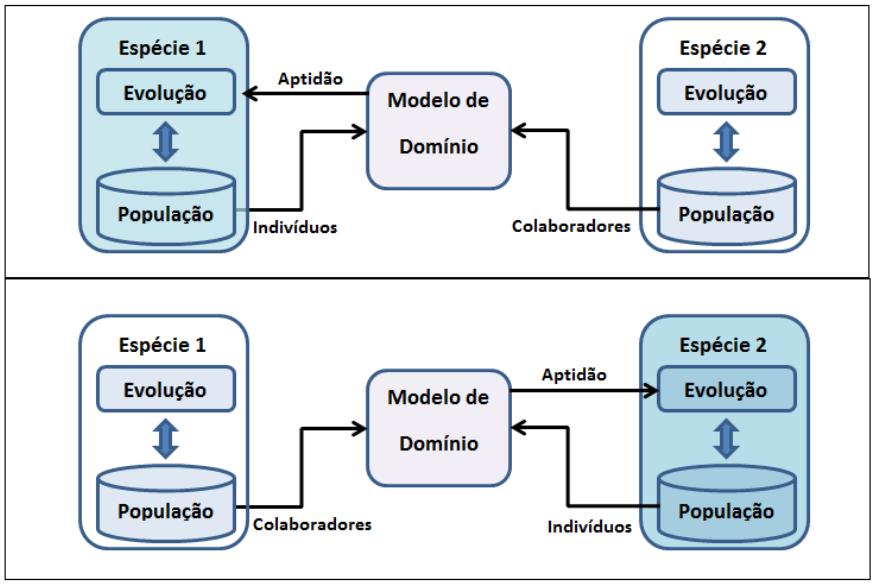

Figura 5: Modelo co-evolucionário cooperativo.

Fonte: Potter e Jong (2000).

A co-evolução cooperativa tende a fazer uso da decomposição do problema para atribuir cada subproblema a uma população. Os colaboradores são os indivíduos que serão selecionados de acordo com sua aptidão para a solução completa do problema. Os indivíduos são avaliados e selecionados em cada geração. Assim, cada população terá seu conjunto de colaboradores que poderão ser selecionados para aderirem a solução completa do problema (POTTER e JONG, 2000).

A escolha de colaboradores é feita segundo três critérios: a pressão da seleção dos colaboradores que é definida pelo peso da avaliação de cada indivíduo em colaborar para a solução do problema, a quantidade de colaboradores que é dada pelo número de colaboradores por população a ser utilizado na avaliação e a associação de avaliação de colaboradores que é a forma de associar um valor de aptidão a um colaborador (WIEGAND *et al*., 2001).

A pressão por seleção dos colaboradores recorre a diversos métodos de avaliação que variam de acordo com o grau de pressão na escolha. Por exemplo: selecionar sempre o melhor indivíduo da geração ou selecionar dois indivíduos de cada geração em que o melhor indivíduo é escolhido e um outro é escolhido de forma

aleatória. Podem ser selecionados dois indivíduos de cada população tal como o melhor e o pior individuo de cada população (POPOVICI *et al*., 2012). A quantidade de colaborados é um fator muito importante e que deve ser muito bem avaliado, pois quanto maior a quantidade dos colaborados, mais eficientes serão os resultados. Porém, um maior número de colaboradores pode ocasionar o efeito colateral de aumento do custo computacional (POTTER e JONG, 2000). A associação de avaliação de colaboradores pode ser determinada como otimista e pessimista. A avaliação otimista é associada a melhor colaboração feita pelo indivíduo. A avaliação pessimista é associada ao valor médio da colaboração do indivíduo (ANGELES, 2010).

Na evolução competitiva a competição entre espécies segue o padrão da natureza em que duas ou mais populações diferentes entram em disputa por um mesmo recurso no meio. Ocorre disputa por alimento, água, locais de reprodução, dentre outros. Essa relação é não harmônica onde os membros de uma espécie evoluem em detrimento dos membros de outra espécie. Esta relação traz prejuízo pelo menos para uma das espécies envolvidas.

Algoritmos genéticos co-evolutivos competitivos imitam os comportamentos da natureza na busca por solução de problemas complexos. A evolução é feita individualmente considerando a influência que uma população exerce nos indivíduos da outra população em evolução. Na dinâmica evolutiva uma população evolui de acordo com a redução da capacidade que a outra população tem para a solução do problema. Por se tratar de uma competição, uma das duas espécies tem maiores prejuízos de acordo com o problema em questão. O pseudocódigo 2.1 encontrado em Figueredo (2004), apresenta um algoritmo co-evolutivo competitivo típico.

Algoritmo 2.1: Pseudocódigo de um Algoritmo co-evolutivo competitivo típico.

```
1. Início
2.    Inicializa População 1 e População 2
3.    Avalia à População 1 em relação à População 2
4.    Avalia à População 2 em relação à População 1
5.   repita
6.        repita
7.            Evolui à População 1
8.            Avalia à População 1 em relação à População 2
9.         até que a condição de parada 1 é satisfeita
10.        repita
11.           Evolui à População 2
12.           Avalia à População 2 em relação à População 1
13.        até que condição de parada 2 é satisfeita
14.       até que condição de parada 3 é satisfeita
15. fim algoritmo
```

O pseudocódigo 2.1 ilustra o comportamento de um algoritmo co-evolutivo competitivo padrão com duas espécies. As Populações 1 e 2 são inicializadas com números aleatórios na Linha 1. Nas Linhas 3 e 4 é calculada a função de avaliação (aptidão) para as Populações 1 e 2. As Populações 1 e 2 são avaliadas uma em relação a outra. Na Linha 7 a População 1 é evoluída, sua função de avaliação é calculada e a População 1 é avaliada em relação a População 2 na Linha 8, iterativamente até que a condição de parada 1 é satisfeita. Na Linha 11 a População 2 é evoluída, sua função de avaliação é calculada e a População 2 é avaliada em relação a População 1 na Linha 12, iterativamente até que a condição 2 é satisfeita. Os blocos, da Linha 6 até 13, de evolução da População 1 e avaliação da População 1 em relação a População 2 e evolução da População 2 e avaliação da População 2 em relação a População 1 é feito iterativamente até que a condição parada 3 é satisfeita. Cada iteração é feita até que a aptidão seja satisfatória em cada caso.

A técnica co-evolutiva foi usada ao problema de classificação de dados a nível interpopulacional onde as populações cooperaram em um regime semi-isolado e

também a nível intra-populacional, onde os classificadores candidatos evoluíram competitivamente com as amostras de treinamento. O algoritmo proposto com a técnica co-evoliva teve melhor desempenho do que algoritmos tradicionais que não utilizam a co-evolução (AUGUSTO, 2009). Outra abordagem fez uso da co-evolução para otimização da programação da produção em refinarias de petróleo. O trabalho teve como objetivo desenvolver um modelo evolucionário para otimizar a programação da produção de produtos derivados do petróleo, o modelo desenvolvido teve boa capacidade em gerar soluções viáveis e sem a necessidade de usar métodos de correções (SIMÃO, 2004).

3. Algoritmos de Seleção de Amostras e Seleção de Variáveis

3.1 Algoritmo Kennard-Stone (K-S)

O algoritmo K-S tenta assegurar que o conjunto de treinamento seja representativo acerca das amostras disponíveis, realizando escolhas de amostras distantes uma das outras no conjunto total das amostras. O algoritmo utiliza a distância Euclidiana d**x**(*p,q*) entre os vetores **x** (resposta instrumental) de cada par (*p;q*) de amostras selecionadas, como pode ser visto na Equação (3-1).

$$d_{\mathbf{x}}(\mathrm{p},\mathrm{q}) = \sqrt{\sum_{j=1}^{n}[\mathbf{x}_p\,(j) - \mathbf{x}_q(j)]^2}; \quad p,q \in [1,N] \qquad (3\text{-}1)$$

Os dados espectrais, **x***p*(j) e **x***q*(j) são as respostas instrumentais no j-ésimo comprimento de onda para as amostras *p* e *q*, respectivamente, e j corresponde aos comprimentos de onda (FILHO, 2007), a seleção inicial é feita entre o par (*p1, p2*) de amostras com maior distância. Este procedimento se repete até que o número de amostras desejado seja alcançado. Na Figura 6 é apresentado um exemplo do critério de seleção de amostras pelo algoritmo K-S.

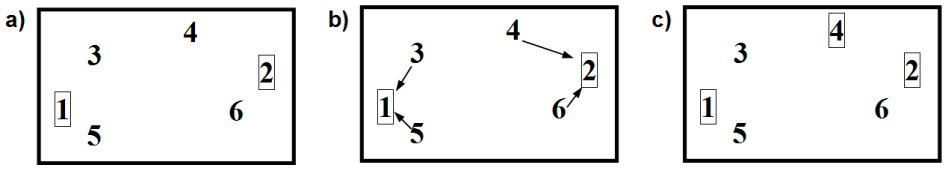

Figura 6: Exemplo de seleção de amostras com algoritmo K-S.

A Figura 6, representa 3 amostras sendo selecionadas pelo algoritmo K-S em um conjunto de 6 amostras. Inicialmente no item a) são selecionas as amostras mais distantes do espaço amostral, neste caso as amostras 1 e 2. Em seguida no item b) é feito o calculo da distância euclidiana entre as amostras não selecionadas, indicada na Figura 6 pelas setas. Armazenam as distâncias menores em relação às amostras selecionadas, o K-S calcula a distância da amostra 3 em relação à amostra 1, a distância da amostra 5 em relação à amostra 1. Essas seriam as distâncias mínimas entre a amostra selecionada 1 e as não selecionadas mais próximas. O mesmo procedimento é realizado com as outras amostras, calcula a distância da amostra 4 em relação à amostra 2 e a distância da amostra 6 em relação à amostra 2. Depois da realização do cálculo mínimo da distância no item c) seleciona a amostra com maior distância em relação às amostras selecionadas, nesse caso, a amostra 4 é selecionada. Esse procedimento se repete determinando a distância de todas as amostras, e o procedimento de seleção de amostras se encerra quando o número de amostra desejada seja alcançado.

O algoritmo K-S segue as seguintes fases a) seleção da amostra mais próxima/ afastada do valor médio das amostras, b) compara a distância entre as amostras do conjunto com as amostras já selecionadas, c) seleção das amostras com maior distância em relação as amostras selecionadas, retornando ao item b) até que a quantidade desejada de amostras seja alcançada (FACCHIN, 2005).

3.2 Algoritmo SPXy

Outra alternativa para a divisão das amostras em calibração e validação, seria o algoritmo proposto por Galvão *et al*. (2005). Denominado SPXY, é uma extensão do algoritmo K-S tem como objetivo, atribuir um aumento na distância atribuída na equação (3-1), levando em consideração a diferença da distância de **x** (variável

independentes), assim também, como a diferença da distância de **y** (variável dependentes), conforme a Equação (3-2).

$$\mathrm{d_y(p,q)} = \sqrt{(\mathbf{y_p} - \mathbf{y_q})}^2 = \left|\mathbf{y_p} - \mathbf{y_q}\right|; \quad \mathrm{p,q} \in [1, \mathrm{N}] \tag{3-2}$$

Contribuindo com igual importância nas distribuições das amostras no espaço (**x,y**), considerando assim, as distâncias d**x**(*p,q*) e d**y**(*p,q*), a distância normalizada é calculada como:

$$d_{\boldsymbol{xy}}(p,q) = \frac{d_{\boldsymbol{x}}(p,q)}{max_{p,q\,\epsilon\,[1,N]}\, d_{\boldsymbol{x}}(p,q)} + \frac{d_{\boldsymbol{y}}(p,q)}{max_{p,q\,\epsilon\,[1,N]}\, d_{\boldsymbol{x}}(p,q)} =; \quad p,q \in [1,N] \tag{3-3}$$

3.3 Gerador de números aleatórios (RNG)

Os números aleatórios são definidos como uma sequência de números em que o próximo número da sequência não se prevê (CASSIA-MOURA *et al*., 2005). Esses números são de extrema importância para uma diversidade de áreas como física nuclear, tomada de decisão, probabilidade, otimização de sistemas, dentre outras. Este trabalho utilizou como fator comparativo os resultados do RNG sendo utilizado para seleção de amostras.

Os computadores digitais são utilizados para gerar números aleatórios, porém esses números gerados por computadores não são considerados totalmente aleatórios, e sim uma sequência de números matematicamente calculada e prevista seguindo uma regra fixada (ROSA *et al.*, 2002). Com intenção de amenizar esses problemas foram criados geradores de números aleatórios baseados em formulas matemáticas e com um valor inicial definido como semente. Um gerador de números aleatórios ideal deve seguir uma distribuição uniforme, os números gerados devem ser

independentes entre si, ou seja, um número gerado não pode interferir no próximo valor gerado, a sequência dos números aleatórios não pode se repetir, e os números devem ser gerados de forma rápida, poupando tempo e recursos computacionais (GRAYBEAL e POOCH, 1980).

O RNG em computadores digitais é atribuído em um espaço S, que evolui em uma sequência finita de números (PAPOULIS *et al*., 1991), como pode ser visto na Equação (3-4).

$$S_{n=} \int (S_n - 1),\ n \geq 1 \tag{3-4}$$

Em que, S_0 é a semente, $\int : S \rightarrow$ S é a função de transição no n-ésimo passo, a saída é dada por $U_n = g\ (S_n)$, sendo $g : S \rightarrow [0,1]$ a função de saída.

3.4 Seleção de Variáveis

As técnicas de reconhecimento de padrões e calibração multivariada quando aplicadas a conjunto de dados com números relativamente alto de variáveis, ficam limitadas. Uma vez que algumas dessas variáveis podem portar a mesma informação (VASCONCELOS, 2011). A eficiência do modelo de calibração está em identificar as variáveis que contenham as informações mais relevantes à solução do problema de interesse, deixando o modelo mais robusto, tendo simplicidade de entendimento e ainda considerando um menor erro de predição.

A utilização e o desempenho de certos métodos de reconhecimento de padrões e de regressão podem ser prejudicados, pois muitas variáveis são irrelevantes ou redundantes (NAES e MEVIK, 2001).

Diante de tais melhorias para o modelo, muitas técnicas têm sido desenvolvidas na busca por essas informações. O que normalmente difere uma

técnica da outra são os procedimentos realizados para selecionar a região em que estão essas informações. Serão apresentas a seguir algumas técnicas que podem apresentar estas informações.

3.5 Mínimos Quadrados Parciais (PLS)

Esta técnica foi desenvolvida por Herman Wold *et al.* (2001) entre os anos de 1975 a 1982. É utilizada principalmente quando o usuário não definiu ou não pode definir o comprimento de onda que devem ser utilizados para a calibração (SANTOS *et al.,* 2005). Usa regressão linear multivariada, relacionando os dados espectrais **X** com as propriedades físicas e químicas de interesse **y** desta forma o PLS procura um conjunto de componentes, fazendo a decomposição simultânea em **X** e **y**, explicando o máximo possível da covariância entre **X** e **y** (ABDI, 2003). Gerando as variáveis latentes através de combinações lineares nas variáveis originais, e assim, transformando os dados em um novo domínio em que não possuem colinearidades entre si. Sugerindo ainda o uso da técnica de validação cruzada para definir o número de componentes a serem retirados (WOLD *et al*., 2001).

A escolha do número de componentes que continuarão no modelo é feita através da avaliação de significância em relação à predição de cada componente, a inclusão de componentes é interrompida à medida que deixam de ser significativos (WOLD *et al*., 2001).

Bueno (2004) comparou os desempenhos das técnicas de PLS e de redes neurais artificiais para caracterização de petróleo e o melhor desempenho foi obtido pela calibração PLS. Ferrer *et al*. (2008) descreve que poucas ferramentas de análise estatística possuem a versatilidade da regressão PLS, oferecendo aplicações de diferentes áreas, como a discriminação e classificação de observações, avaliação de desempenho do comportamento humano e ciências sociais.

O grande desafio e limitação desta técnica esta na dificuldade de interpretação, uma vez que houve uma transformação nas variáveis de entrada para a construção das variáveis latentes e que a regressão é realizada neste novo domínio dos dados transformado, não há possibilidades de uma interpretação direta nos dados.

3.6 Algoritmo das Projeções Sucessivas (APS)

Proposto em 2001 por Araújo e colaboradores (SHAMSIPUR *et al*., 2006) tendo como objetivo selecionar variáveis para a construção de modelos multivariados. É uma técnica de seleção em que a variável é incorporada em cada interação, atribuindo assim, uma nova variável com a menor multicolinearidade possível em relação às variáveis que já foram selecionadas (PEREIRA e PASQUINI, 2010). De uma forma geral, a seleção de variáveis APS utiliza três conjuntos de dados: treinamento, validação e teste. É feito uma medida para avaliar se os valores previstos a partir das medidas de **X** são condizentes com os de **y** (SOARES *et al*., 2013). O APS é dividido em três fases. A fase inicial é gerado o conjunto de variáveis com menor redundância, ou seja, minimizando a multicolinearidade, usando apenas as informações da matriz **X**. Na fase 2, é selecionado o subconjunto que tiver a melhor habilidade de previsão, realiza-se uma avaliação no conjunto de variáveis geradas na fase inicial e que estão correlacionadas com as respostas de interesse. E obtém o melhor resultado da habilidade de previsão através da Raiz Quadrada do Erro Médio Quadrático de Validação (RMSEP) (ARAÚJO *et al*., 2001), descrito na Equação (3-5).

$$RMSEP = \sqrt{\frac{1}{K_v} \sum_{K=1}^{Kv} (y_v^K - \hat{y}_v^K)^2} \qquad (3\text{-}5)$$

Sendo: *Kv* o número de amostras do conjunto de validação, y_v^K *e* $\hat{y}_v^K$ são os valores de referência e os valores de previsão do parâmetro de interesse na *k*-ésima amostra de validação.

A fase 3 tem por objetivo, eliminar do modelo final as variáveis que não contribuem significativamente com a capacidade preditiva do modelo por meio de um teste estatístico (PRESS) (GALVAO *et al*., 2008).

Fernandes *et al.* (2011) usou o algoritmo de K-S para a seleção do conjunto de treinamento e teste e o algoritmo SPA na calibração multivariada na quantificação de biodiesel, Sanches e colaboradores fizeram uso na determinação de dipirona em ampolas fechadas.

As técnicas de seleção de variáveis estão relacionadas com o princípio de que uma pequena quantidade de variáveis é suficiente para gerar bons preditores (SOARES *et al*., 2013). No entanto, o sucesso destas técnicas está diretamente ligado à escolha adequada das amostras utilizadas no conjunto de treinamento.

4. Resultados e Discussões

Atualmente algoritmos evolutivos simples têm sido usados para a solução de problemas em calibração multivariada (NAES *et al*., 2002). O propósito desse trabalho é desenvolver um algoritmo co-evolutivo para a seleção de amostras e variáveis de modo a considerar a influência de uma sobre a outra no sentido de melhorar a capacidade preditiva do modelo em comparação com métodos tradicionais que realizam essas etapas em separado. O algoritmo proposto é denominado Algoritmo Genético Co-evulutivo Proposto (AGCP). O AGCP é igualmente avaliado a partir do valor do erro de predição do modelo criado e em relação ao número de variáveis selecionadas para compor o modelo.

A Figura 7 representa os dois módulos e as diferentes fases do AGCP. O módulo da esquerda faz a seleção de amostras e o módulo da direita faz a seleção de variáveis. As fases em cada módulo são de um algoritmo evolutivo padrão. Após a seleção das amostras no módulo da esquerda, o conjunto de amostras selecionadas é enviado ao modulo responsável pela seleção de variáveis, em que funcionará como colaboradores a solução do problema. Co-evolutivamente busca as melhores amostras e variáveis frente a solução do problema como um todo com taxa de mutação que varia de zero a 5%. O AGCP foi desenvolvido no ambiente Matlab versão 7.11 e os testes feito em um computador com processador Intel Core i3-330 (2,13 GHz) com 4 Gb de memória RAM.

A seguir são apresentados os resultados da seleção de amostras empregando algoritmos tradicionais RNG, K-S e SPXy, São apresentados e discutidos os resultados dos algoritmos tradicionais para a seleção de amostras associados a algoritmos de seleção de variáveis. De forma aleatória foram criadas as junções do algoritmo RNG e SPA, K-S e SPA, SPXy e SPA, RNG e PLS, K-S e PLS, SPXy e PLS, K-S e algoritmo genético simples (*Simple Genetic Algorithm*, SGA) e

consequentemente será apresentado o desempenho do algoritmo Genético co-evolutivo (AGCP).

Os valores de RMPSEP mostrados na Tabela 1 são obtidos com o emprego do RNG. O RMSEP obtido com o K-S e SPXy são 2,8270 e 1,4567, respectivamente (SANTIAGO, 2013). Os resultados da Tabela 1 mostra que o algoritmo RNG obtém o menor RMSEP em relação aos algoritmos K-S e SPXy. Por outro lado, o RNG obtém o maior RMSEP e fica em média superior quando comparado com os outros algoritmos. O melhor RMSEP é encontrado com o algoritmo SPXy que tem o melhor desempenho entre os algoritmos de seleção de amostras testados.

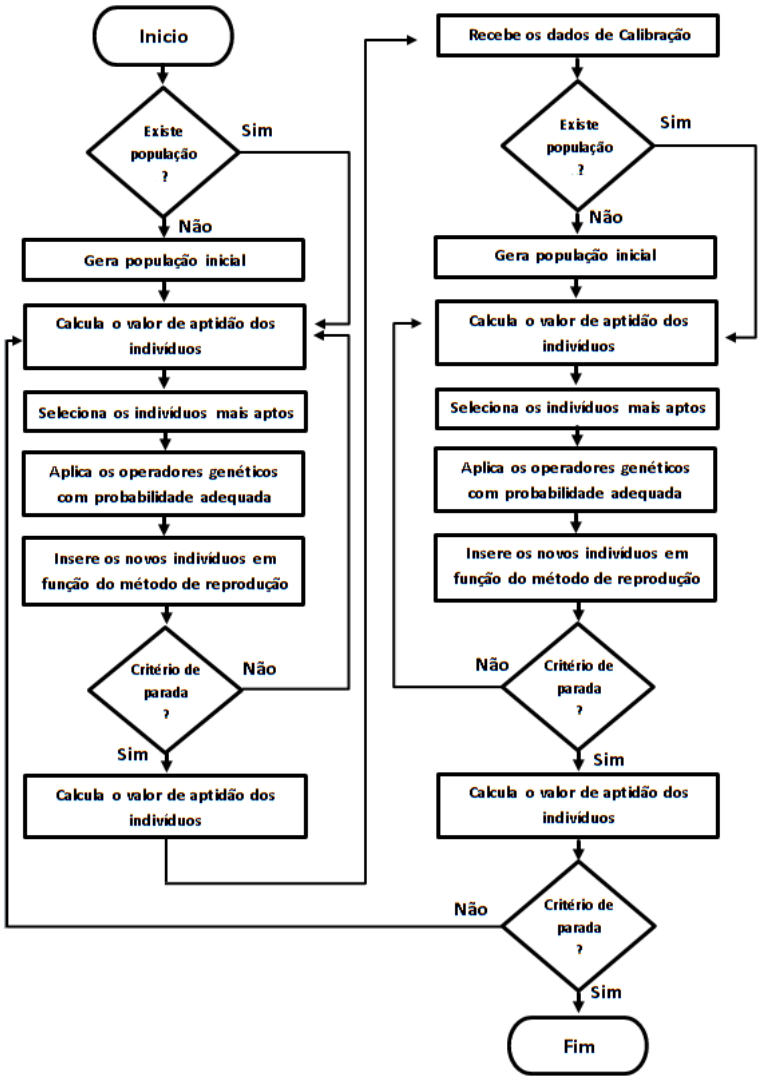

Figura 7: Algoritmo genético co-evolutivo proposto (AGCP).

Tabela 1: Resultados do algoritmo RNG na seleção de amostras.

	RNG
Menor RMSEP	0,8893
Maior RMSEP	5,7882
Desvio padrão	1,1918

Fonte: Santiago (2013).

Os algoritmos RNG e SPA foram executados em conjunto, onde os seus resultados são observados através da Tabela 2. Os dois algoritmos foram utilizados para a seleção de amostras e também para a seleção de variáveis. Obtiveram uma média de 24 variáveis selecionadas, ficando com uma variação de no mínimo 10 e no máximo 50 variáveis. Os algoritmos K-S e SPA foram executados de forma conjunta. Assim também como o algoritmo SPXy e SPA, também para a seleção de amostras e seleção de variáveis, conforme os resultados que podem ser observados na Tabela 3.

Tabela 2: Resultados dos algoritmos RNG e SPA na seleção de

	RNG-SPA
Menor RMSEP	0,2171
Maior RMSEP	0,2517
Desvio padrão	0,0097

Fonte: Santiago (2013).

Tabela 3: Resultados dos algoritmos K-S com SPA e SPXy com

	K-S-SPA	SPXy-SPA
Variáveis	38	22
RMSEP	0,2491	0,2368

Fonte: Santiago (2013).

De acordo com os resultados apresentados na Tabela 2 e Tabela 3, é possível observar que todos os algoritmos que foram utilizados em conjunto com o algoritmo de projeções sucessivas (SPA), para a seleção de amostras e também para a seleção de

variáveis, tiveram melhorias significativas em relação a seção anterior, onde não havia a execução conjunta, mostrando a grande necessidade em realizar a seleção de variáveis na construção de modelos de calibração multivariada. Os algoritmos RNG e PLS, sendo executados juntos, têm seus resultados apresentados na Tabela 4, o qual apresentou melhores resultados com 22 variáveis latentes. Os resultados de RMSEP dos algoritmos K-S com PLS e SPXy com PLS para a seleção de amostras e variáveis, podem ser vistos na Tabela 5, esses algoritmos foram executados com número de variáveis latentes de 1 a 30. Os resultados mostram que os algoritmos tiveram bons desempenhos em comparação com os resultados anteriores que utilizaram o algoritmo SPA.

Tabela 4: Resultados dos algoritmos RNG e PLS na seleção de amostras e variáveis.

	RNG-PLS
Variáveis latentes	22
RMSEP	0.1777
Média RMSEP	0,2070
Desvio padrão	0,0113

Fonte: Santiago (2013).

Os resultados do desempenho do algoritmo genético simples, sendo executado em conjunto com os algoritmos K-S e SPXy são apresentados na Tabela 6.

Tabela 5: Resultados dos algoritmos KS com PLS e SPXY com

	K-S-PLS	SPXy-PLS
Variáveis latentes	14	20
RMSEP	0,2071	0,1973

Fonte: Santiago (2013).

Tabela 6: Resultados dos algoritmos K-S com AGS e SPXy com

	K-S-AGS	SPXy-AGS
Média de variáveis	123	115
Maior RMSEP	0,4332	0,4065
Menor RMSEP	0,2064	0,2348
Desvio padrão RMSEP	0,0162	0,0195

Fonte: Santiago (2013).

Os dados de desempenho dos algoritmos K-S com o AGS e também do SPXy com o AGS, são muito parecidos o que retrata bom desempenho em relação aos resultados dos outros algoritmos apresentados, mostrando eficiência e concorrência com os algoritmos tradicionais.

4.1 Algoritmo Genético Co-evolutivo (AGCP)

O desempenho do AGCP para a seleção de amostras e a seleção de variáveis pode ser visto na Tabela 7, onde também estão sendo apresentados os parâmetros que foram utilizados na execução do algoritmo, como taxa de mutação, tamanho da população, quantidade de gerações, quantidade média das amostras selecionadas e quantidade média das variáveis selecionadas.

Tabela 7: Resultados do algoritmo Genético Co-evolutivo Proposto na seleção

	AGCP
Taxa de Mutação	5%
Tamanho da população	50
Quantidade de gerações	100
Amostras (Média)	292
Variáveis (Média)	63
Maior RMSEP	0,0701
Menor RMSEP	0,0470
Média RMSEP	0,0581
Desvio padrão RMSEP	0,0060

Fonte: Santiago (2013).

Os resultados do desempenho do AGCP na seleção de amostras e seleção de variáveis mostrados na Tabela 7 demonstra que o AGCP é uma técnica competitiva. Os resultados obtidos para a seleção de amostra e seleção de variáveis é superior em relação às técnicas tradicionais testadas. Na comparação com o AGS, que obteve melhor desempenho quando executado em conjunto com o SPXy, o AGCP tem um erro de predição 4 vezes menor. O número de variáveis selecionadas pelo AGCP é mais da metade das variáveis selecionadas pelo SGA. A Figura 8 apresenta o desempenho da evolução das amostras durante a execução do algoritmo AGCP. Mostra que no decorrer das iterações sempre houve melhoras até a estabilização da evolução das amostras.

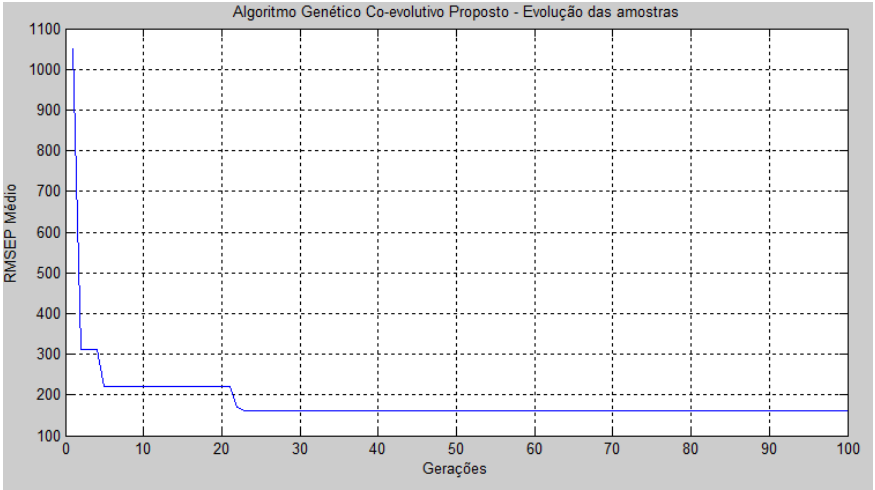

Figura 8: Evolução do algoritmo genético co-evolutivo proposto.

A figura 8 apresenta a evolução das amostras, permitindo visualizar uma melhoria nos seus valores durante as gerações. Nota-se após a geração 20 uma estabilidade nos seus valores, deixando claro que o número de gerações foi suficiente para a convergência do algoritmo. A Figura 9 apresenta a evolução do algoritmo, no caso, o algoritmo executa um total de cem iterações, mostrando uma melhora do erro médio quadrático durante as iterações, nota-se ainda, que entre as iterações 90 e 100,

quase não tivemos melhorias, o que confirma que a quantidade de gerações foram suficientes para a convergência do algoritmo.

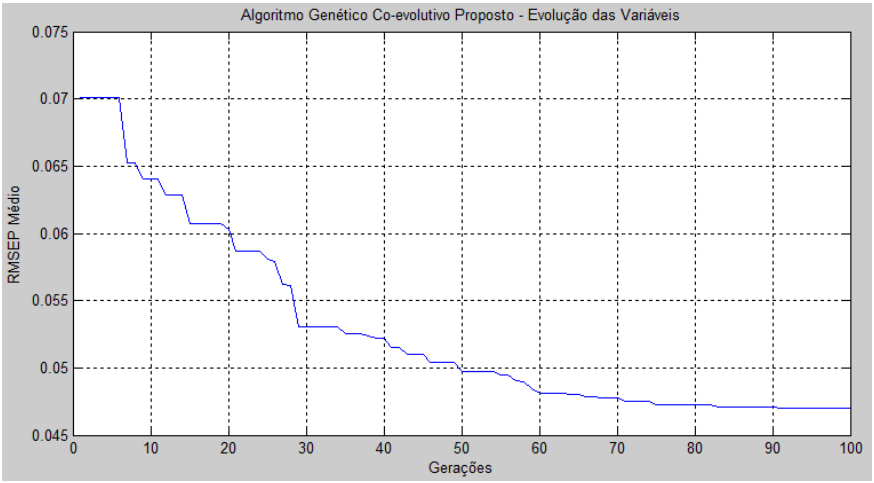

Figura 9: Evolução do algoritmo genético co-evolutivo proposto.

Se for observado apenas os gráficos das Figuras 8 e 9, que representam o comportamento da população de amostras e da população de variáveis durante o processo de evolução, conclui-se que há uma evolução gradativa, não havendo em nenhum momento uma situação de perda nas evoluções, permitindo excluir a possibilidade da existência de uma relação competitiva (desarmônica), uma vez que esta relação traz prejuízos para uma ou para ambas as populações. Assim, temos um ambiente cooperativo entre as populações, o que não permite concluir qual o tipo de cooperação foi estabelecida, sabendo que as relações harmônicas (cooperativas) trazem benefícios as duas ou pelo menos uma das populações. Essas relações podem ser: Protocooperação (em que ambas as populações têm benefícios), Inquilinismo (em que uma população tem benefício e a outra não sofre alterações), Comensalismo (em que uma população tem benefício e a outra não sofre alterações) ou Mutualismo (em que ambas as populações têm benefícios). No entanto, para fazer tais afirmações, seria necessário um embasamento teórico mais aprofundado e realizações de testes direcionados a este fim.

A Tabela 8 apresenta os dados resultantes do desempenho de cada algoritmo testado, podendo ser visualizado os melhores desempenhos das técnicas, com base nas diferenças do erro médio quadrático de cada resultado.

Tabela 8: Comparativo dos valores RMSEP de todos os algoritmos

	RMSEP
RNG	2,5936
K-S	2,8270
SPXy	1,4567
RNG-SPA	0,2373
K-S-SPA	0,2491
SPXy-SPA	0,2368
RNG-PLS	0,2070
K-S-PLS	0,2071
SPXy-PLS	0,1973
KS-AGS	0,2422
SPXy-AGS	0,2348
AGCP	0,0581

A comparação de valores mostrada na Tabela 8, o AGCP tem um desempenho de pelo menos 70% em relação as técnicas tradicionais testadas, descreve o bom desempenho do AGCP frente as outras técnicas tradicionais, mostrando eficiência e robustez, principalmente com base no erro de predição, no processo de seleção de amostras e também na seleção de variáveis.

5. Conclusões

Os algoritmos tradicionais utilizados em problemas de calibração multivariada fazem uso de um algoritmo que faça o pré-processamento dos dados, e outro que estabeleça o modelo. Desta forma, os dois algoritmos trabalham de forma independentes, desconsiderando os resultados causados de um ao outro. Tendo em vista essa deficiência foi proposto um único algoritmo co-evolutivo que realizasse essas duas tarefas, tanto a seleção de amostra como também a seleção de variáveis. Foi estabelecido um estudo de caso que teve como base, dados obtidos por espectroscopia de infravermelho próximo, com o objetivo de verificar a concentração de proteínas nas amostras de trigos.

A análise do desempenho do AGCP foi feita com a utilização de resultados de algoritmos tradicionais, sendo executados em conjunto, um que fizesse a seleção de amostras e outro a seleção de variáveis. Foi possível observar que a técnica proposta teve o melhor desempenho entre todos os resultados, o desempenho teve como fator observado o menor erro de predição.

É importante salientar que os resultados provam a importância da pré-seleção de dados levando em consideração os resultados que serão causados ao processo de seleção de variáveis, e da mesma forma o contrário, isso foi um grande diferencial que trouxe o sucesso nos resultados do algoritmo. A técnica se mostrou muito competitiva e inovadora ao processo de calibração multivariada.

Referências Bibliográficas

ABDI, Hervé. Multivariate analysis. **Encyclopedia for research methods for the social sciences. Thousand Oaks: Sage**, p. 699-702, 2003.

ANGELES, Mary Stankovich. **Use of dynamic pool size to regulate selection pressure in cooperative coevolutionary algorithms**. Nova Southeastern University, 2010.

ANZANELLO, Michel José. Seleção de variáveis com vistas à classificação de bateladas de produção em duas classes. **Gestão & Produção**, v. 16, n. 4, p. 526-533, 2009.

ARAÚJO, Mário César Ugulino *et al.* The successive projections algorithm for variable selection in spectroscopic multicomponent analysis.**Chemometrics and Intelligent Laboratory Systems**, v. 57, n. 2, p. 65-73, 2001.

AUGUSTO, Douglas Adriano. **PROGRAMAÇÃO GENÉTICA MULTI-POPULACIONAL E CO-EVOLUCIONÁRIA PARA CLASSIFICAÇÃO DE DADOS**. 2009. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

BREMERMANN, Hans J.; ROGSON, M.; SALAFF, S. Global properties of evolution processes. **Natural automata and useful simulations**, p. 3-41, 1966.

BROWNE, Michael W.; CUDECK, Robert. Single sample cross-validation indices for covariance structures. **Multivariate Behavioral Research**, v. 24, n. 4, p. 445-455, 1989.

BUENO, A. F. **Caracterização de petróleo por espectroscopia no infravermelho próximo**. 2004. Tese de Doutorado. Tese de Mestrado, Universidade estadual de Campinas.

CASSIA-MOURA, R. *et al.* Yet another application of the Monte Carlo method for modeling in the field of biomedicine. **Computer methods and programs in biomedicine**, v. 78, n. 3, p. 223-235, 2005.

CHAUCHARD, F. *et al.* Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. **Chemometrics and Intelligent Laboratory Systems**, v. 71, n. 2, p. 141-150, 2004.

COUTURE, Raymond; L'ECUYER, Pierre. On the lattice structure of certain linear congruential sequences related to AWC/SWB generators. **Mathematics of Computation**, v. 62, n. 206, p. 799-808, 1994.

DA ROSA, Fernando Henrique Ferraz Pereira; JUNIOR, Vagner Aparecido Pedro; COLLI, Eduardo. Gerando números aleatórios. **Laboratório de Matemática Aplicada Prof. Dr. Eduardo Colli**, 2002.

DANTAS FILHO, H. A. Desenvolvimento de técnicas quimiométricas de compressão de dados e de redução de ruído instrumental aplicadas a óleo diesel e madeira de eucalipto usando espectroscopia NIR. 2007.

DARWIN, Charles. On the origins of species by means of natural selection.**London: Murray**, p. 247, 1859.

EHRLICH, Paul R.; RAVEN, Peter H. Butterflies and plants: a study in coevolution. **Evolution**, p. 586-608, 1964.

ELIASSON, Ann-Charlotte; LARSSON, Kåre. **Cereals in breadmaking: a molecular colloidal approach**. New York: Marcel Dekker, 1993.

FACCHIN, S. **Técnicas de análise multivariável aplicadas ao desenvolvimento de analisadores virtuais.** M.Sc, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2005.

FERNANDES, David Douglas Sousa *et al.* Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. **Talanta**, v. 87, p. 30-34, 2011.

FERREIRA, Márcia MC *et al.* Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, p. 724-731, 1999.

FERRER, Alberto *et al.* PLS: A versatile tool for industrial process improvement and optimization. **Applied Stochastic Models in Business and Industry**, v. 24, n. 6, p. 551-567, 2008.

FIGUEREDO, Grazziela Patrocinio. **Algoritmos Genéticos na Simulaçao da Evoluçao das Bibliotecas de Genes do Sistema Imune**. 2004. Tese de Doutorado. Master's Thesis, Universidade Federal do Rio de Janeiro-COPPE/UFRJ.

FRASER, Alex S. Simulation of genetic systems by automatic digital computers vi. epistasis. **Australian Journal of Biological Sciences**, v. 13, n. 2, p. 150-162, 1960.

GALVAO, Roberto Kawakami Harrop *et al.* A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. **Chemometrics and intelligent laboratory systems**, v. 92, n. 1, p. 83-91, 2008.

GALVAO, Roberto Kawakami Harrop *et al.* A method for calibration and validation subset partitioning. **Talanta**, v. 67, n. 4, p. 736-740, 2005.

GASARCH, William I. Gurari Eitan. An introduction to the theory of computation. Principles of computer science series. Computer Science Press, Rockville, Md., 1989, xii+ 314 pp. **The Journal of Symbolic Logic**, v. 56, n. 01, p. 338-339, 1991.

GOLDBERG, D. E. Genetic algorithms in search, optimization, and machine learning, addison-wesley, reading, ma, 1989. **K. Ohno, K. Esfarjani, and Y Kawazoe: Computational Materi**.

GOLDBERG, David. The Design of Innovation (Genetic Algorithms and Evolutionary Computation). 2002.

GOLDBERG, David E. **The design of innovation: Lessons from and for competent genetic algorithms**. Springer Science & Business Media, 2013.

GRAYBEAL, Wayne T.; POOCH, Udo W. **Simulation: principles and methods**. Cambridge, MA: Winthrop Publishers, 1980.

GREFENSTETTE, John J. *et al.* AAAI **Spring Symposium on Adaptation, Coevolution, and Learning in Multiagent Systems**. 1996.

HO L LA ND, J. H. **Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence,** 1975.

IORIO, Antony W.; LI, Xiaodong. A cooperative coevolutionary multiobjective algorithm using non-dominated sorting. In: **Genetic and Evolutionary Computation–GECCO 2004**. Springer Berlin Heidelberg, 2004. p. 537-548.

KENNARD, Ronald W.; STONE, Larry A. Computer aided design of experiments. **Technometrics**, v. 11, n. 1, p. 137-148, 1969.

LASZTITY, Radomir. **The chemistry of cereal proteins**. CRC Press, 1995.

LEARDI, Riccardo. Application of genetic algorithm-PLS for feature selection in spectral data sets. **Journal of Chemometrics**, v. 14, n. 5-6, p. 643-655, 2000.

LI, Heng *et al.* The sequence alignment/map format and SAMtools.**Bioinformatics**, v. 25, n. 16, p. 2078-2079, 2009.

LINDEN, Ricardo. **Algoritmos genéticos (2a ediçao)**. Brasport, 2008.

LUCASIUS, Carlos B.; KATEMAN, Gerrit. Understanding and using genetic algorithms Part 1. Concepts, properties and context. **Chemometrics and intelligent laboratory systems**, v. 19, n. 1, p. 1-33, 1993.

MARTENS, Harald; MARTENS, Magni. **Multivariate analysis of quality: an introduction**. John Wiley & Sons, 2001.

MICHALEWICZ, Zbigniew; SCHOENAUER, Marc. Evolutionary algorithms for constrained parameter optimization problems. **Evolutionary computation**, v. 4, n. 1, p. 1-32, 1996.

MITCHELL, Melanie. **An introduction to genetic algorithms**. MIT press, 1998.

NAES, Tormod *et al.* **A user friendly guide to multivariate calibration and classification**. NIR publications, 2002.

NÆS, T.; MEVIK, B.-H. **Understanding the collinearity problem in regression and discriminant analysis**. Journal of Chemometrics, 15(4):413–426, 2001.

NEVES, Ana Carolina de Oliveira. **Espectroscopia no infravermelho próximo e métodos de calibração multivariada aplicados à determinação simultânea de parâmetros bioquímicos em plasma sanguíneo**. 2013.

OLIVEIRA, Flavia CC *et al.* A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos. **Química Nova**, v. 27, n. 2, p. 218-225, 2004.

PAPOULIS, A. “Probability, Random Variables, and Stochastic Processes”, Third Edition, McGraw-Hill, 1991.

PASQUINI, Celio. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198-219, 2003.

PEŇA-REYES, CarlosAndrés; SIPPER, Moshe. Fuzzy CoCo: A cooperative-coevolutionary approach to fuzzy modeling. **Fuzzy Systems, IEEE Transactions on**, v. 9, n. 5, p. 727-737, 2001.

PEREIRA, Claudete Fernandes; PASQUINI, Celio. A flow system for generation of concentration perturbation in two-dimensional correlation near-infrared spectroscopy: application to variable selection in multivariate calibration. **Applied spectroscopy**, v. 64, n. 5, p. 507-513, 2010.

POPOVICI, Elena *et al.* Coevolutionary principles. In: **Handbook of Natural Computing**. Springer Berlin Heidelberg, 2012. p. 987-1033.

POTTER, Mitchell A.; DE JONG, Kenneth A. Cooperative coevolution: An architecture for evolving coadapted subcomponents. **Evolutionary computation**, v. 8, n. 1, p. 1-29, 2000.

ROSIN, Christopher D.; BELEW, Richard K. New methods for competitive coevolution. **Evolutionary Computation**, v. 5, n. 1, p. 1-29, 1997.

SANTIAGO, Kelton de Sousa. Algoritmo evolutivo de cromossomo duplo para calibração multivariada. 2013.

SANTOS, Alexandre F. *et al.* Monitoring and control of polymerization reactors using NIR spectroscopy. **Polymer-Plastics Technology and Engineering**, v. 44, n. 1, p. 1-61, 2005.

SANTOS, D. A. *et al.* Calibração multivariada multiproduto e espectroscopia ultravioleta na determinação da acidez total em bebidas industrializadas a base de soja e néctar de frutas. **Blucher Chemical Engineering Proceedings**, v. 1, n. 2, p. 3417-3423, 2015.

SHAMSIPUR, Mojtaba *et al.* Ant colony optimisation: a powerful tool for wavelength selection. **Journal of Chemometrics**, v. 20, n. 3-4, p. 146-157, 2006.

SIMÃO, Leonardo Mendes. **Otimização da Programação da Produção em Refinarias de Petróleo utilizando Algoritmos Genéticos e Co-evolução Cooperativa**. 2004. Tese de Doutorado. PUC-Rio.

SKOOG, Douglas A.; WEST, Donald M.; HOLLER, F. James. **Fundamentos de química analítica**. Reverté, 1997.

SOARES, Sófacles Figueredo Carreiro *et al.* The successive projections algorithm. **TrAC Trends in Analytical Chemistry**, v. 42, p. 84-98, 2013.

SOUSA, Leonardo Chagas de. Espectroscopia na região do infravermelho próximo para predição de características da madeira para produção de celulose. 2008.

SRINIVAS, Mandavilli; PATNAIK, Lalit M. Adaptive probabilities of crossover and mutation in genetic algorithms. **Systems, Man and Cybernetics, IEEE Transactions on**, v. 24, n. 4, p. 656-667, 1994.

SUMATHI, Sai; HAMSAPRIYA, T.; SUREKHA, P. **Evolutionary intelligence: an introduction to theory and applications with Matlab**. Springer Science & Business Media, 2008.

VASCONCELOS, F. V. C. **Uso da região espectral de sobretons para determinação do teor de biodiesel e classificação de misturas diesel/biodiesel adulteradas com óleo vegetal**. Master's thesis, Universidade Federal da Paraíba, 2011.

VON ZUBEN, Fernando José. **Projeto Automático de Sistemas Nebulosos: Uma Abordagem Co-Evolutiva**. 2002. Tese de Doutorado. FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO Projeto Automático de Sistemas Nebulosos: Uma Abordagem Co-Evolutiva Myriam Regattieri De Biase da Silva Delgado Prof. Dr. Fernando José Von Zuben (Orientador) Prof. Dr. Fernando Gomide (Co-orientador) Tese apresentada à Pós-graduação da Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como requisito parcial à obtenção do grau de Doutor em Engenharia Elétrica na área de Engenha-ria de Computação. Banca examinadora: Profa. Dra. Marley Maria Bernardes Rebuzzi Vellasco–DEE, PUC-Rio.

WIEGAND, R. Paul; LILES, William C.; DE JONG, Kenneth A. An empirical analysis of collaboration methods in cooperative coevolutionary algorithms. In: **Proceedings of the genetic and evolutionary computation conference (GECCO)**. 2001. p. 1235-1245.

WOLD, Svante. Chemometrics; what do we mean with it, and what do we want from it?. **Chemometrics and Intelligent Laboratory Systems**, v. 30, n. 1, p. 109-115, 1995.

WOLD, Svante *et al.* Some recent developments in PLS modeling.**Chemometrics and intelligent laboratory systems**, v. 58, n. 2, p. 131-150, 2001.

YAGER, Ronald R.; FILEV, Dimitar P. Essentials of fuzzy modeling and control. **New York**, 1994.