Seleção de bandas espectrais apoiada pela meta-heurística PSO para predição do teor de alumínio de amostras de solo

Arion de Campos Jr[1], **José Carlos F. da Rocha**[1], **Giancarlo Rodrigues**[1]

[1]Programa de Pós-Graduação em Computação Aplicada - UEPG
Ponta Grossa – PR – Brasil

Abstract. *ERD-IR is a technique that can be employed to make a model of soil nutrient prediction by correlating sample data to the respective reference value obtained by chemical analysis. Such data are organized as attributes of a set of high-dimensional records and making a model from these data involves difficulties that impair its performance. To overcome such difficulties, the use of evolutionary algorithms for Feature Selection has shown to be promising. The aim of this article is to identify, with Particle Swarm Optimization meta-heuristic, the relevant wavelengths for predicting the aluminum content of soil samples from the Campos Gerais region. Results suggest that, for this scenario, few iterations and small swarm size provide the best subsets.*

Resumo. *ERD-IR é uma técnica que pode ser utilizada para criar um modelo de predição do teor de nutrientes do solo ao correlacionar os dados de amostras ao respectivo valor de referência obtido por análise química. Tais dados são dispostos como atributos de um conjunto de registros de alta dimensionalidade e elaborar um modelo a partir desses dados envolve dificuldades que prejudicam seu desempenho. Para superar tais dificuldades, a utilização de algoritmos evolucionários para Seleção de Atributos tem se mostrado promissora. O objetivo deste artigo é identificar, com o apoio da meta-heurística de Otimização por Enxame de Partículas, os comprimentos de onda relevantes à predição do teor de alumínio de amostras de solo da região dos Campos Gerais. Resultados sugerem que, para este cenário, poucas iterações e tamanho de enxame reduzido fornecem os melhores subconjuntos.*

1. Introdução

A espectroscopia de refletância difusa do infravermelho (ERD-IR) é um método em química analítica que se utiliza da fração de energia refletida (informação espectral) ao incidir um nível de energia conhecido sobre uma amostra para identificar e quantificar sua composição. Em decorrência de seu custo acessível, de não utilizar reagentes e de não danificar a amostra avaliada, a ERD-IR tem se mostrado uma alternativa viável para a estimativa do teor dos nutrientes do solo. Uma atividade essencial para o manejo do solo.

A estimativa das quantidades de nutrientes a partir de dados de EDR-IR demanda um modelo de predição que associe a informação espectral da amostra ao respectivo valor de referência. Uma abordagem para o desenvolvimento de tais modelos é o emprego da Aprendizagem Máquina (AM). Nesta abordagem, cada comprimento de onda do espectro amostral é um atributo do conjunto de dados usado no treinamento dos modelo. Contudo, a alta dimensionalidade dos conjuntos de dados gerados pela EDR-IR pode levar ao sobreajuste dos modelos.

As técnicas de seleção automática de atributos (FSS - *Feature Subsection Selection*) visam determinar um subconjunto de variáveis preditoras que seja efetivo para a tarefa alvo. Isto reduz a dimensionalidade dos dados e reduz o risco de sobreajuste. O alto custo no processamento da FSS baseada em *wrappers* tem motivado o emprego de métodos de busca evolucionária na sua realização [Viniski and Guimarães 2017]. Considerando o exposto, este trabalho avalia o desempenho do algoritmo de Otimização por Enxame de Partículas (PSO, do inglês *Particle Swarm Optimization*) na identificação de atributos relevantes para estimativa do teor de alumínio, o qual em grandes concentrações prejudicam o desenvolvimento das culturas.

O PSO foi concebido para solucionar problemas contínuos, porém a seleção de atributos é um problema discreto. Uma forma de contornar esta dificuldade é estabelecer um limiar que determina se os atributos serão ou não selecionados. Como o limiar interfere nos resultados do PSO, ele precisa ser ajustado para a tarefa alvo. Em vista disso, este trabalho incorporou um esquema de limiarização ao procedimento de busca. O efeito da configuração dos parâmetros de busca do PSO sobre o procedimento de regressão também foi analisada. Resultados sugerem que esta abordagem reduz significativamente a dimensionalidade da base de dados, reduzindo o risco que sobreajuste ao mesmo tempo que possibilita definição de um modelo de regressão cujo desempenho preditivo que se mostrou superior àquele descrito em trabalhos correlatos nos experimentos realizados. Devido a limitações de espaço, uma síntese comparativa entre os trabalhos correlatos foi disponibilizada em *https://www.deinfo.uepg.br/∼arion/SBIAGRO2023/SBIAGRO23-Tab2.pdf*.

2. Referencial Teórico

O solo é um sistema quimicamente complexo composto de água, ar e matéria orgânica e inorgânica e são primordiais para a agricultura. Um dos nutrientes presentes no solo é o Alumínio (Al), o qual em grandes concentrações compromete a absorção da água e nutrientes do solo e assim prejudica o desenvolvimento das culturas [Fageria et al. 1988]. Portanto, detectar a concentração de Al no solo é importante para determinar se alguma prática de manejo será necessária.

Métodos convencionais de amostragem do solo utilizam-se de análises químicas, as quais exigem tempo para serem executadas e empregam reagentes químicos custosos que implicam no descarte da amostra após a análise [Rossel et al. 2006]. Neste contexto, a espectroscopia de refletância difusa do infravermelho apresenta-se como uma alternativa capaz de realizar rapidamente a amostragem simultânea de diversos atributos sem as desvantagens dos métodos convencionais [Viniski and Guimarães 2017]. Tal técnica incide uma quantidade de energia conhecido sob a amostra e mensura a quantidade de energia refletida. Como cada molécula possui uma capacidade de absorção específica em relação ao nível de energia fornecido, é possível identificar e quantificar a composição química

das amostras através da interpretação dos valores de refletância (comprimentos de onda) registrados no seu espectro.

A estimativa do teor dos nutrientes do solo pela ERD-IR é feita a partir de um modelo de predição e o desempenho de um modelo é reconhecido através de seu Coeficiente de Determinação (R^2) e Raiz do Erro Quadrático Médio (RMSE), os quais são medidas descritivas da qualidade do seu ajuste e da sua acurácia, respectivamente.

Devido à quantidade de comprimentos de onda, é necessário extrair matematicamente a informação do espectro e correlacioná-la ao atributo desejado para que a espectroscopia possa ser utilizada na estimativa do teor dos nutrientes do solo [Rossel et al. 2006]. À medida que a dimensão do conjunto aumenta cada vez mais instâncias de treinamento são exigidas para manter o bom desempenho do algoritmo de aprendizado [Russell and Norvig 2010]. Uma maneira de contornar tais complicações é através da redução da dimensionalidade dos conjuntos.

A FSS é uma técnica de pré-processamento para redução da dimensionalidade de conjuntos de dados que procura pelo subconjunto mínimo de seus atributos que forneça o melhor desempenho de predição possível quando utilizado na elaboração de um modelo [Tan 2010]. Para isso, atributos com ruído, redundantes ou irrelevantes que poderiam reduzir a exatidão do modelo são detectados e removidos. Na FSS baseada em *wrappers*, a remoção dos atributos irrelevantes é realizada por um procedimento iterativo que testa diferentes conjuntos de atributos e seleciona aquele que atende a critérios de precisão previamente estabelecidos. Uma das desvantagens desta abordagem é seu alto custo computacional.

O PSO [Kennedy and Eberhart 1995] é uma meta-heurística evolucionária que simula o comportamento de um bando de pássaros[1] para resolução dos problemas de busca e otimização em que é aplicado. Nesse algoritmo, a quantidade de dimensões d é igual ao número de variáveis a serem otimizadas no problema. Para buscar pela solução ótima, a cada iteração do algoritmo as partículas são reposicionadas de acordo com as melhores posições obtidas por si e suas partículas vizinhas até o momento. Ao final da execução do algoritmo, a partícula que detém o melhor posicionamento/aptidão é definida como a solução ótima do problema. Além dos parâmetros de controle utilizados para a atualização da velocidade, a quantidade de partículas do enxame e número de iterações também interferem no desempenho e devem ser ajustados empiricamente.

Considerando que cada algoritmo de AM adota uma abordagem específica para propor um modelo, diversos trabalhos de ERD-IR aplicados à estimativa do teor de nutrientes de amostras do solo foram propostos [Rossel et al. 2006]. A FSS, por sua vez, mostrou-se adequada para melhorar o desempenho de predição dos modelos de ERD-IR, pois favorece a operação dos algoritmos de AM [Viniski and Guimarães 2017]. Trabalhos correlatos comumente têm como objetivo propor melhorias ao algoritmo que favoreçam a FSS [Nguyen et al. 2017] ou abordagens híbridas que realizam a FSS ao mesmo tempo em que parâmetros de controle do algoritmo de AM são otimizados. Outrossim, quase a totalidade desses trabalhos tratam de tarefas de classificação e não regressão dos dados.

[1]No bando, um pássaro (partícula) líder conduz o restante dos pássaros até uma fonte de alimento com determinada velocidade. Se outro pássaro identifica uma fonte com potencial superior à do líder atual, este assume a liderança e passa a conduzir o bando para a nova direção

De maneira geral, tais publicações não investigaram a aplicação dessas técnicas sobre conjuntos de dados de ERD-IR para estimativa do teor de alumínio. Ademais, nenhum deles considerou amostras de solo da região dos Campos Gerais.

3. Material e Métodos

Foram utilizados dois conjuntos de dados de ERD-IR: um para executar a FSS e elaborar o modelo de predição – *Conjunto de dados 1* – e outro para validar seu resultado – *Conjunto de dados 2* – conforme sugerido por [Tan 2010]. Cada conjunto contém dados de amostras coletadas em propriedades agrícolas localizadas na região dos Campos Gerais (Piraí do Sul e Ponta Grossa, respectivamente), a qual está entre 24°e 26°sul de latitude, 49°e 51°oeste de longitude com altitude entre 600 e 1.300 metros do nível do mar.

As amostras coletadas em ambas as propriedades foram encaminhadas ao laboratório de análises físico químicas da Fundação ABC[2], empresa que atua no desenvolvimento de pesquisa aplicada à agricultura, para que as medições fossem realizadas. Um espectrômetro foi utilizado para mensurar a refletância difusa das amostras de ambos os conjuntos de dados. O equipamento realiza leituras no intervalo de 400 a 2.500 nm com um intervalo de 2 nm entre os comprimentos de onda, gerando dados de refletância de 1.050 comprimentos de onda. Os valores registrados nos conjuntos de dados, por sua vez, são da absorbância aparente de cada comprimento de onda, o qual é resultado da conversão de $log(1/R)$ onde R representa o respectivo valor de refletância.

A análise de referência foi executada após a análise espectral, já que utiliza reagentes químicos que poluem as amostras e implicam no seu descarte. Os resultados dessa análise então anexados aos respectivos dados espectrais, o que finalizou a composição das instâncias dos conjunto de dados. Cada conjunto de dados foi constituído por 1.050 atributos oriundos da análise espectral mais o atributo meta obtido pela análise de referência, totalizando 1.051 atributos. Entretanto, na presença dos 1.050 atributos de entrada não foi possível elaborar sequer um modelo de predição a partir do algoritmo de Regressão Linear Múltipla (RLM), portanto tais conjuntos fornecem o cenário ideal para investigar a aplicabilidade da técnica de FSS evolucionária com PSO.

Para executar a FSS com o PSO é necessário elaborá-la como um problema de otimização que atenda à representação manipulada pelo algoritmo. As partículas do enxame foram codificadas de forma que cada índice do vetor posição correspondeu ao índice de um atributo de entrada do conjunto de dados, portanto a dimensão das partículas, d, foi igual a 1.050. Em virtude da FSS ser um problema do tipo discreto, para utilizar a versão contínua do PSO é necessário estabelecer um valor de limiar que determina se o valor contínuo existente no índice da partícula indica a seleção ou não do respectivo atributo do conjunto de dados.

Tran et al. [Tran et al. 2016] identificaram que o valor do limiar interfere na quantidade de atributos selecionados e que o valor ideal varia conforme o conjunto de dados. Assim, o valor que proporcionou os melhores subconjuntos foi investigado. Os valores candidatos foram os mesmos utilizados por esses autores: $0,05$; $0,2$; $0,4$; $0,6$; $0,8$ e $0,95$.

Para estabelecer a função de aptidão do algoritmo de otimização, quatro etapas são definidas: ***Etapa 1:*** *são identificados os índices do vetor de posição da partícula que*

[2]http://fundacaoabc.org/

apresentam valor igual ou superior ao limiar pré-estabelecido; ***Etapa 2:*** *uma equação de regressão com os respectivos atributos selecionados é elaborada.* ***Etapa 3:*** *a equação é executada e obtém um modelo de regressão junto com seus respectivos indicadores de desempenho (detalhados a seguir). Cada partícula do enxame propõe um modelo de regressão candidato;* ***Etapa 4:*** *por fim, o valor referente ao potencial da solução fornecida pela partícula é retornado, o que encerra sua avaliação.*

Em relação aos indicadores de desempenho, o Critério de Informação de Akaike (AIC, do inglês *Akaike's Information Criterium*) [Akaike 1998] permite avaliar quão bem um modelo ajusta-se aos dados levando em consideração sua capacidade preditiva (através da sua Máxima Verossimilhança ou RMSE) e sua complexidade (número de atributos utilizados). Dentre os modelos avaliados, aquele com menor valor de AIC representa o melhor modelo aproximado [Symonds and Moussalli 2011].

O cálculo do AIC na função de aptidão utilizou o valor de RMSE do modelo elaborado pelo algoritmo de regressão [3]. A Equação 1 apresenta como o AIC foi calculado, na qual n é o número de amostras utilizadas para treinamento, ln é a operação de logaritmo e m a quantidade de atributos do modelo. O primeiro termo dessa equação trabalha com a capacidade preditiva do modelo enquanto o segundo penaliza sua complexidade.

$$AIC = n \cdot \ln(RMSE) + 2 \cdot m \tag{1}$$

A versão *SPSO2011* [Clerc 2012] foi utilizada neste artigo por ser considerada uma versão padrão. Ao utilizar o PSO para resolução de um problema de otimização é necessário estabelecer seus parâmetros. Foram avaliadas diversas combinações de número de iterações, tamanho do enxame e valores de limiar a fim de identificar aquela que forneceu o melhor resultado de FSS no *Conjunto de dados 1*: **Dimensão do espaço de busca (d):** 1.050; **Limites do espaço de busca:** 0,0 $\sim$ 1,0; **Critério de parada:** Atingir o número máximo de iterações (40, 70, 100); **Tamanho do enxame:** 20, 40, 60, 80, 100.

Cada combinação foi repetida 30 vezes utilizando uma semente distinta para geração de números aleatórios, a qual foi modificada de forma controlada. Esse procedimento foi necessário devido à natureza estocástica do PSO e para reprodutibilidade dos resultados. O critério de parada do algoritmo foi atingir o número máximo de iterações.

Para reconhecer o melhor subconjunto de atributos dentre os 30 disponíveis na combinação ideal identificada, cada um foi validado no *Conjunto de dados 2* e então o melhor foi apontado segundo os mesmos critérios da identificação da combinação ideal. Esse foi o procedimento de seleção adicional apontado por Xue et al. [Xue et al. 2016].

As combinações de parâmetros e os resultados da validação foram comparados estatisticamente através do Teste de Friedman com 5% de significância para investigar a existência de diferença estatística entre os mesmos. Se a diferença é constatada, então o teste *post-hoc* de Friedman é utilizado para identificar quais pares apresentavam essa característica. Por meio desses testes e da avaliação empírica dos resultados, as melhores combinações e o melhor subconjunto para predição do teor de Alumínio trocável dos conjuntos manipulados são identificados.

[3] Disponível no software R, o qual procura identificar um relacionamento linear entre os atributos de entrada e o atributo meta, o qual não poderia ser executado na presença dos 1.050 atributos.

4. Resultados e Discussão

A combinação ótima do número de iterações, tamanho do enxame e valor de limiar foi investigada a fim de identificar a configuração de parâmetros que proporcionou a seleção dos atributos que geraram o modelo de regressão com os maiores índices de predição do teor de alumínio. No problema proposto uma solução (global ou local) deveria otimizar o ajuste do modelo e penalizar a alta dimensionalidade. Isto é obtido pelo índice AIC. Os resultados mostram que os menores de AIC, alcançados com o limiar $0,95$, estavam associados a modelos com apenas um atributo (comportamento esperado, em função do elevado limiar). No entanto, tais modelos mostraram-se inadequados em termos de R^2 e RMSE [Symonds and Moussalli 2011]. Considerando que o AIC não mostrou-se relacionado aos demais indicadores de ajuste, os valores de R^2 e RMSE foram testados na sequência.

Nas situações em que o Teste de Friedman constatou diferença estatística para os valores de R^2 ou RMSE o teste *post-hoc* foi utilizado para apontá-la entre os pareamentos. Esse teste foi conduzido primeiramente sobre os valores de R^2, mas nas situações em que os pareamentos não apresentaram diferença estatística quanto a esta métrica também foi conduzido sobre os valores de RMSE para identificar uma única solução.

A primeira característica exposta pelos diferentes valores de limiar remete à quantidade média de atributos selecionados, pois, de maneira geral, quanto menor seu valor mais atributos foram selecionados. Independente do número de iterações ao utilizar 20, 40 ou 60 partículas no enxame o limiar $0,6$ favoreceu a seleção dos atributos que produziram os modelos com melhor desempenho, enquanto que com 80 e 100 partículas o limiar $0,4$ logrou tal feito. O tamanho de enxame ideal foi identificado a partir dos valores de R^2 e RMSE dos modelos produzidos pelos limiares ideais de cada um, os quais foram comparados entre si para identificar o tamanho ideal a ser utilizado em cada número de iterações.

O Teste de Friedman constatou diferença estatística entre os valores de R^2 e RMSE dos modelos, logo o teste *post-hoc* foi aplicado. No entanto, o resultado das avaliações e do teste *post-hoc* foram idênticos para cada número de iterações e tamanho de enxame avaliados. Assim, o tamanho de enxame 20 foi selecionado devido ao menor custo computacional em decorrência da menor quantidade de partículas avaliadas. O teste de Friedman também não detectou diferença estatística entre os valores de R^2 e RMSE para os modelos gerados com 40, 70 e 100 iterações. Com base neste resultado, escolheu-se a configuração com 40 iterações.

A combinação ideal de parâmetros para executar a FSS com PSO no *Conjunto de Dados 1* foi a seguinte: 40 iterações, 20 partículas no enxame e limiar em $0,6$. Considerando esta configuração, após trinta rodadas, os melhores resultados são apresentados na Tabela 1[4]. O experimento evidencia a drástica redução de atributos do conjunto de dados.

Em função do modelo com melhor R^2 no treinamento, a Equação 2 apresenta a fórmula de regressão utilizada por este modelo para predição do teor de alumínio, na qual $12,88$ é seu valor de interceptação ou ajuste. Variáveis iniciadas pela letra r referem-se à

[4]Resultados com todas as combinações avaliadas estão disponíveis em https://www.deinfo.uepg.br/∼arion/SBIAGRO2023/SBIAGRO23-Tab1.pdf

Tabela 1. Desempenho de treinamento e validação do melhor modelo gerado através da combinação de 40 iterações, 20 partículas e limiar 0,6

		Treinamento			Validação		
Repetição	Nº Atributos	AIC	R^2	RMSE	AIC	R^2	RMSE
10	**22**	**46,05**	**0,61**	**1,24**	**88,61**	**0,50**	**5,96**

refletância da amostra no respectivo comprimento de onda[5] e valores que antecedem-nas são os coeficientes atribuídos.

$$\begin{aligned} TeorAluminio = {} & 12,88 + 404,56 * r486 - 940,07 * r554 + 4.821,80 * r594 \\ & - 6.002,88 * r612 + 1.911,41 * r678 - 12.964,86 * r1008 \\ & + 17.804,04 * r1020 - 5.996,58 * r1050 + 2.164,77 * r1152 \\ & + 2.980,35 * r1272 - 5.124,90 * r1276 + 1.180,70 * r1404 \\ & - 1.605,05 * r1430 + 3.095,31 * r1626 - 3.112,30 * r1864 \\ & + 1.236,11 * r2040 + 963,61 * r2086 + 1.592,58 * r2274 \\ & - 755,02 * r2308 - 4.507,30 * r2332 + 2.636,73 * r2368 \\ & + 218,62 * r2460 \end{aligned} \tag{2}$$

Em relação às pesquisas publicadas na área e considerando os resultados obtidos, o melhor modelo identificado obteve desempenho de predição superior àqueles de Rossel et al. [Rossel et al. 2006], elaborado na região espectral do MIR, e de Terra, Demattê e Rossel [Terra et al. 2015], elaborado na região do Vis-NIR, com dados transformados e com muito mais instâncias de treinamento disponíveis. Tais autores não utilizaram a FSS para elaborar seus modelos, portanto o potencial da FSS fica evidenciado.

5. Conclusão

Neste trabalho, uma técnica de FSS evolucionária foi aplicada sobre um conjunto de dados de espectroscopia para selecionar um subconjunto de atributos relevantes à predição do teor de alumínio de amostras de solo. A FSS foi abordada como um problema de otimização e seu objetivo foi minimizar o valor médio de AIC na elaboração dos modelos dos subconjuntos candidatos pelo algoritmo de RLM. Em função de seu desempenho na FSS, a versão contínua do PSO foi utilizada e exigiu investigação da melhor configuração de parâmetros possível.

Tal investigação indicou 40 iterações, 20 partículas no enxame e limiar $0,6$ como a combinação ideal, a qual além de fornecer os modelos de melhor desempenho utilizou o menor número de iterações e de tamanho de enxame analisados. Os modelos obtidos através dessa combinação evidenciam o potencial do método adotado, já que o modelo mais complexo utilizou menos de 10% do total de atributos do conjunto original, enquanto o modelo com melhor desempenho utilizou 22 atributos de entrada.

Os resultados obtidos pela aplicação da FSS evolucionária foram positivos. Sem a redução da dimensionalidade não seria possível obter sequer um modelo gerado pelo algoritmo de RLM na presença dos 1.050 atributos de entrada do conjunto de dados. Enquanto que pela sua utilização apenas 22 atributos foram necessários para explicar 50% da variação dos dados ($R^2 = 0,5$) com este algoritmo. Essa pequena quantidade de atributos selecionados comparada à quantidade original do conjunto evidencia o potencial

[5]Exemplo: r486 refere-se ao valor de refletância no comprimento de onda de 486 nm.

da modelagem adotada para realizar a FSS evolucionária com algoritmo PSO, portanto esta é uma técnica oportuna e indicada para FSS em conjuntos de dados de espectroscopia.

Considerando que o coeficiente de determinação obtido foi baixo, algumas oportunidades de trabalhos futuros foram identificadas, as quais podem aprimorar os resultados deste trabalho. Destacamos a adoção de novos métodos de inicialização das partículas e tratar esta pesquisa como um problema multi-objetivo, minimizando o número de atributos e maximizando a capacidade preditiva, simultaneamente.

Referências

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics)*, pages 199–213.

Clerc, M. (2012). Standard Particle Swarm Optimisation.

Fageria, N. K., Ballgar, V. C., and Wright, R. J. (1988). Aluminum toxicity in crop plants. *Journal of Plant Nutrition*, 11(3):303–319.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE.

Nguyen, H. B., Xue, B., Andreae, P., and Zhang, M. (2017). Particle Swarm Optimisation with genetic operators for feature selection. In *2017 IEEE Congress on Evolutionary Computation (CEC)*, pages 286–293. IEEE.

Rossel, R. V., Walvoort, D., McBratney, A., Janik, L., and Skjemstad, J. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131(1-2):59–75.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition.

Symonds, M. R. E. and Moussalli, A. (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*, 65(1):13–21.

Tan, K. H. (2010). *Principles of soil chemistry*. CRC press, 4 edition.

Terra, F. S., Demattê, J. A., and Viscarra Rossel, R. A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geoderma*, 255-256:81–93.

Tran, C. T., Zhang, M., Andreae, P., and Xue, B. (2016). Improving performance for classification with incomplete data using wrapper-based feature selection. *Evolutionary Intelligence*, 9(3):81–94.

Viniski, A. D. and Guimarães, A. M. (2017). Técnicas de seleção de atributos para mineração de dados de alta dimensionalidade gerados por espectroscopia no infravermelho próximo–NIR. In *Anais SULCOMP*, volume 8, Criciuma, SC.

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2016). A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation*, 20(4):606–626.