CADERNO DE CIÊNCIAS AGRÁRIAS
*Agrarian Sciences Journal*

ICA INSTITUTO DE CIÊNCIAS AGRÁRIAS

UFMG

# Técnica de seleção de variáveis em problemas de classificação de falhas aplicada em processo industrial usando o algoritmo genético MOEADD

**Leonardo Macedo Freire[1], Luiz Carlos Gabriel Filho[2]*, Luana Michelly Aparecida da Costa[3], Marcos Flávio Silveira Vasconcelos D'Angelo[4], Maurílio José Inácio[5], Rosivaldo Antônio Gonçalves[6]**

**Resumo**

Neste trabalho é proposto um método de seleção de variáveis denominado MOEADD-KNN-M, que é baseado no algoritmo genético MOEADD (Evolutionary Many-Objective Optimization Algorithm Based on Dominance and Decomposition), no algoritmo de classificação KNN (K-nearest neighbors), e em operadores genéticos adaptados. A abordagem adotada no algoritmo proposto é bi-objetivo, onde um objetivo é minimizar a quantidade de variáveis da solução e outro objetivo é minimizar a taxa de erro de classificação de falhas. Foram realizados experimentos com o método proposto empregando dados de um processo industrial petroquímico real, denominado Tennessee Eastman para classificação de falhas, e os resultados obtidos foram comparados com outros algoritmos. Os resultados demonstraram que o método proposto leva a soluções com baixo erro de classificação e pouca quantidade de sensores, que são as quantidades procuradas para serem minimizadas. Sendo assim, essa abordagem se mostrou promissora para a aplicação na seleção de variáveis em problemas de classificação de falhas em processos industriais.

**Palavras Chave:** Indústria. KNN. Operadores Genéticos. Inteligência Computacional.

# Variable selection technique in fault classification problems applied in industrial process using the MOEADD genetic algorithm

**Abstract**

In this work we propose a method of variable selection called MOEADD-KNN-M, which is based on the genetic algorithm MOEADD (Evolutionary Many-Objective Optimization Algorithm Based on Dominance and Decomposition), on the classification algorithm KNN (K-nearest neighbors), and in adapted genetic operators. The approach adopted in the proposed algorithm is bi-objective, where one objective is to minimize the amount of solution variables and another objective is to minimize the failure classification error rate. Experiments were performed with the proposed method using data from a real petrochemical industrial process, called Tennessee Eastman for failure classification, and the results were compared with other algorithms. The results showed that the proposed method leads to solutions with low classification error and low number of sensors, which are the quantities sought to be minimized. Thus, this approach has shown promise for application in the selection of variables in fault classification problems in industrial processes.

**Keywords:** Industry. KNN. Genetic Operators. Computational intelligence.

[1]Universidade Estadual de Montes Claros - Unimontes. Montes Claros, MG. Brasil.
https://orcid.org/0000-0001-9559-0303
[2]Universidade Estadual de Montes Claros - Unimontes/Universidade Federal de Minas Gerais - UFMG-ICA. Montes Claros, MG. Brasil.
https://orcid.org/0000-0003-1740-9753
[3]Universidade Estadual de Montes Claros - Unimontes. Montes Claros, MG. Brasil.
https://orcid.org/0000-0002-9939-274X
[4]Universidade Estadual de Montes Claros - Unimontes. Montes Claros, MG. Brasil.
https://orcid.org/0000-0001-5754-3397
[5]Universidade Estadual de Montes Claros - Unimontes. Montes Claros, MG. Brasil.
https://orcid.org/0000-0003-0744-0845
[6]Universidade Estadual de Montes Claros - Unimontes. Montes Claros, MG. Brasil.
https://orcid.org/0000-0003-3275-5224
*Autor para correspondência: luizcgf@ufmg.br

**Recebido para publicação em 29 de setembro de 2019. Aceito para publicação em 31 de outubro de 2019.**
**e-ISSN: 2447-6218 / ISSN: 2447-6218 /** 

## Introdução

O monitoramento de eventos anormais em processos de tomada de decisão se beneficia do uso de sistemas computacionais, em parte devido à complexidade e o alto volume de dados que devem ser analisados. Por isso, a importância do estudo da automação em atividades desta área, denominada Gerenciamento de Eventos Anormais (Abnormal Event Management, AEM) (Venkatasubramanian *et al.*, 2003c), é de fundamental importância para a Indústria. Porém, automatizar as tarefas de AEM ainda é um grande desafio para a indústria e a comunidade acadêmica.

Como componente das atividades de AEM, há os Sistemas de detecção e Diagnósticos de falhas (Fault Detection and Diagnosis, FDD), que são sistemas responsáveis pelo monitoramento de processos, onde é possível detectar falhas, classificando-as quanto ao seu tipo, diagnosticando suas causas e origens. Na literatura há várias propostas de sistemas de FDD utilizando métodos baseados em modelos quantitativos, modelos qualitativos e métodos baseados em dados históricos do processo (Venkatasubramanian *et al.*, 2003c,a,b). Uma etapa importante relacionada ao diagnóstico em sistemas FDD é a classificação de falhas, onde detectada uma falha pelo sistema, identifica-se as causas da condição anormal. A correta classificação das falhas em tempo hábil permite a correção do problema, sendo fundamental para a segurança do processo e a produtividade.

Um problema enfrentado pelos sistemas de FDD, no entanto, é o fato de que os processos industriais modernos estão se tornando cada vez mais complexos, tanto em níveis estruturais quanto de automação. O grande número de variáveis lidos pelos sensores dispersos pelas plantas podem ser irrelevantes e/ou redundantes. A eliminação destas variáveis pode levar a sistemas de FDD mais eficientes e robustos, e a escolha das variáveis adequadas para o monitoramento, no entanto, não é uma tarefa fácil (Foster *et al.*, 2015). Esse problema é abordado na literatura na área de mineração de dados como seleção de variáveis, e é NP-hard de complexidade O(2n), sendo frequentemente tratado por meio de métodos heurísticos e metaheurísticos, classificados como filter e wrapper (Chandrashekar e Sahin, 2014).

Os métodos de wrapper utilizam métricas de performance do algoritmo de classificação para guiar a seleção de variáveis (Chandrashekar e Sahin, 2014; Guyon e Elisseeff, 2003; Kohavi e John, 1997). Esse processo de seleção de variáveis dos métodos wrapper podem ser determinísticos, usando algoritmos de seleção seqüencial, como o Sequential Forward Selection (SFS) (Whitney, 1971), ou Sequential Backward Selection (SBS) (Marill e Green, 1963). Podem ser também por meio de heurísticas ou metaheurísticas, que são capazes de explorar o espaço de busca global, como os algoritmos evolutivos Differential Evolution (Al-Ani *et al.*, 2013), Ant Colony Optimization (ACO) (Allegrini e Olivieri, 2011), Particle Swarm Optimization (PSO) (Ahmad, 2015), e Simulated Annealing (Brusco, 2014), etc.

Os métodos de filter não utilizam os algoritmos de classificação e fazem a seleção de variáveis como uma etapa de pré-processamento dos dados. Para isso, utilizam características

intrínsecas dos dados selecionando as variáveis baseados em critérios de relevância e/ou similaridade entre elas, como o coeficiente de correlação de Pearson ou Informação Mútua (Yu e Liu, 2004). Nos métodos wrapper as abordagens para a seleção de variáveis podem ser mono-objetivos, onde busca-se melhorar a métrica de performance extraída do algoritmo de classificação para as soluções. Pode ser também multiobjetivo, havendo mais de uma métrica de performance a ser analisada. Ou além da métrica analisada, busca-se explicitamente minimizar a complexidade das soluções (quantidade de variáveis). Na abordagem multiobjetivo é retornando pelo método um conjunto de soluções não dominadas, onde dentre elas pode-se escolher aquela com o melhor custo-benefício (Mukhopadhyay *et al.*, 2013).

Neste trabalho é feita uma proposta de um método wrapper multiobjetivo de seleção de variáveis para aplicação em classificação de falhas baseadas em dados históricos de um processo petroquímico, usando o algoritmo genético multiobjetivo MOEADD (Evolutionary Many-Objective Optimization Algorithm Based on Dominance and Decomposition) proposto em Li *et al.* (2014), e o algoritmo de classificação KNN (K-nearest neighbors) com operadores genéticos adaptados ao problema. No método proposto, denominado MOEADD-KNN-M, o algoritmo MOEADD foi adaptado para atuar como o mecanismo de busca para a seleção de variáveis, devido ao seu bom desempenho em problemas de otimização multiobjetivo (Li *et al.*, 2014).

De uma forma geral, em qualquer processo industrial é necessário ter um controle de cada uma das componentes do processo. Instalar sensores em cada uma das componentes pode gerar custos desnecessários. Uma pergunta a ser feita é: Onde instalar estes sensores? Podemos instalar sensores nas variáveis mais representativas, obtidas do método de seleção de variáveis, sem a necessidade de instalar sensores em todas as máquinas, possibilitando assim uma redução de custos, e por sua vez, melhoria e eficiência na produção.

## Materiais e Métodos

Neste trabalho foi implementado um método wrapper de seleção de variáveis, denominado MOEADD-KNN-M, onde a taxa de erro de classificação é dada pelo KNN, como uma das funções fitness a serem minimizadas. Além disso, foram adaptados operadores genéticos de cruzamento no MOEADD para o problema proposto.

Na literatura existe um banco de dados muito utilizado, chamado *Tennesee Eastman Process* (TEP), que é um processo industrial petroquímico real para avaliar métodos de controle e monitoramento de processos. Estes dados podem ser obtidos no site do *Massachusetts Institute of Technology* (MIT). Neste conjunto de dados existem dados para treinamento e teste, tanto das condições de operação normal do sistema (*Normal Operation Conditions*, NOC), como também algumas falhas em algumas das componentes deste processo. Os dados de treinamento são compostos por 500 instâncias para o NOC e 480 para cada tipo de falha. Os dados de teste são compostos por 960 instâncias para o NOC e cada tipo de falha. As falhas e o NOC são compostos por instâncias de 52 variáveis, que são as componentes do processo petroquímico.

Para cada falha, das 960 instâncias dos dados de teste, foram descartadas as 160 primeiras por apresentarem dados do NOC. Foram descartadas dos dados de treinamento e de testes as falhas dos tipos 3, 9, 15 e 21, porque são falhas de difícil detecção, pois não existem diferenças significativas em relação aos dados do NOC. Dos 17 tipos de falhas, foram utilizadas as 480 instâncias de cada para treinamento e 800 para testes.

**Método proposto: MOEADD-KNN-N**

Neste trabalho o algoritmo genético MOEADD proposto por Li *et al.* (2014) foi adaptado para funcionar como método de seleção de variáveis, usando o KNN como classificador. O MOEADD possui parâmetros próprios, e foram mantidos os mesmos valores propostos por Li *et al.* (2014):

Uma penalidade $\theta=5.0$, usada na abordagem de decomposição, chamada de *penalty-based boundary intersection* (PBI).

Um parâmetro $T=20$, que indica a quantidade de sub-regiões vizinhas a uma sub-região. A vizinhança de uma sub-região é importante, pois há indivíduos associados a essas sub-regiões, sendo que na vizinhança de uma sub-região qualquer será feita a seleção de indivíduos, para serem aplicados os operadores genéticos.

Um parâmetro $\delta=0.9$ sendo a probabilidade dos indivíduos selecionados para a aplicação dos operadores genéticos, na vizinhança de tamanho T, de uma sub-região qualquer. Onde 1-$\delta$ é a probabilidade de serem selecionados considerando toda a população.

No método proposto MOEADD-KNN-M, cada indivíduo representa as variáveis que terão sua instância de dados usada pelo classificador para classificar as falhas do processo. Como o importante para uma menor taxa de erro do classificador é tanto a quantidade de variáveis da solução, quanto quais são essas variáveis selecionadas em si, a representação foi pensada para que os operadores genéticos possam atuar explicitamente sobre essas duas características.

Cada indivíduo é representado por um vetor preenchido pelas variáveis que o compõe, sendo assim, cada vetor tem o tamanho da quantidade de variáveis pertencentes a ele. O tamanho do indivíduo pode variar de 1 a 52, sendo as variáveis representadas por inteiros de 1 a 52. Os indivíduos da população inicial foram gerados aleatoriamente, tanto em relação ao seu tamanho e suas variáveis, e inicializada com uma população de 150 indivíduos.

Os operadores de cruzamento e mutação foram adaptados da seguinte forma: o operador de cruzamento é baseado no *Order crossover* (OX) proposto por Davis (1985), é um operador de representações discretas. Como na representação utilizada. Os vetores que representam os indivíduos podem ter tamanhos diferentes, o operador foi adaptado a essa condição.

O cruzamento pode acontecer segundo uma probabilidade PC e retorna um filho apenas e seu funcionamento é da seguinte forma: Dois pontos de corte são gerados aleatoriamente entre as posições do maior indivíduo. O material genético entre os pontos de corte desse indivíduo então é copiado para o filho. O restante do material genético do filho é preenchido com o do segundo pai a partir do segundo ponto de corte em uma busca cíclica pelas posições que existirem nesse pai depois do segundo ponto de corte e antes do primeiro. A busca garante que o filho não tenha variáveis repetidas. O valor da probabilidade de cruzamento PC usado foi 0.9. Para exemplificar isto, veja a figura 1 abaixo:

A mutação proposta para a representação é dividida em duas partes. A primeira parte procura mudar as variáveis que já existem no indivíduo sem alterar seu tamanho. O indivíduo pode ser submetido a essa mutação com uma probabilidade $\Omega$. Sendo submetido a ela cada variável desse indivíduo pode ser mudada para outra aleatoriamente, e que já não exista nele, segundo uma probabilidade $\omega$.

A segunda parte pode ocorrer com uma probabilidade 1-$\Omega$, ou seja, os indivíduos são obrigatoriamente submetidos à apenas uma parte da mutação. Essa parte busca aumentar ou diminuir o número de variáveis do indivíduo. Aleatoriamente escolhe se uma posição de 1 a 52 (tamanho máximo que um indivíduo pode ter) e se tal número for maior que seu tamanho atual aumenta-o com variáveis escolhidas aleatórias, e que já não existam nele. Se o número escolhido for menor retiram-se variáveis aleatórias dele. O valor de $\Omega$ usado foi 0.7, e o de $\omega$ foi de 0.3. A figura 2 abaixo ilustra os operadores de mutação:

Figura 1 – Etapas do cruzamento

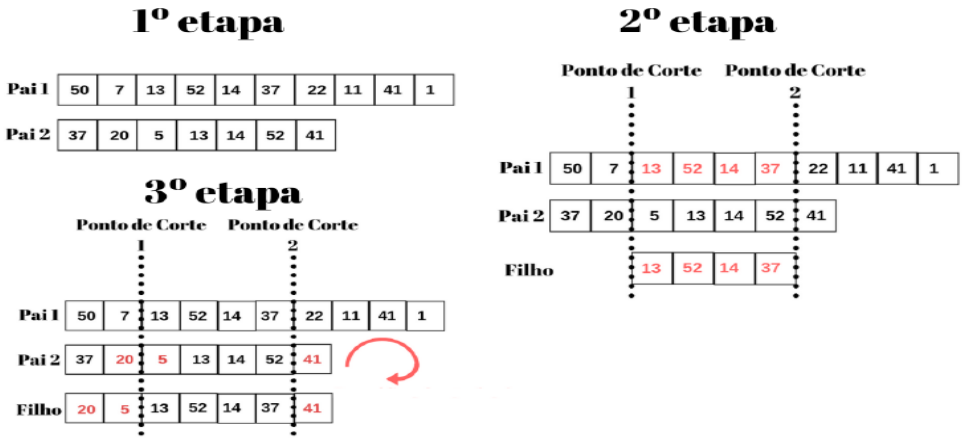
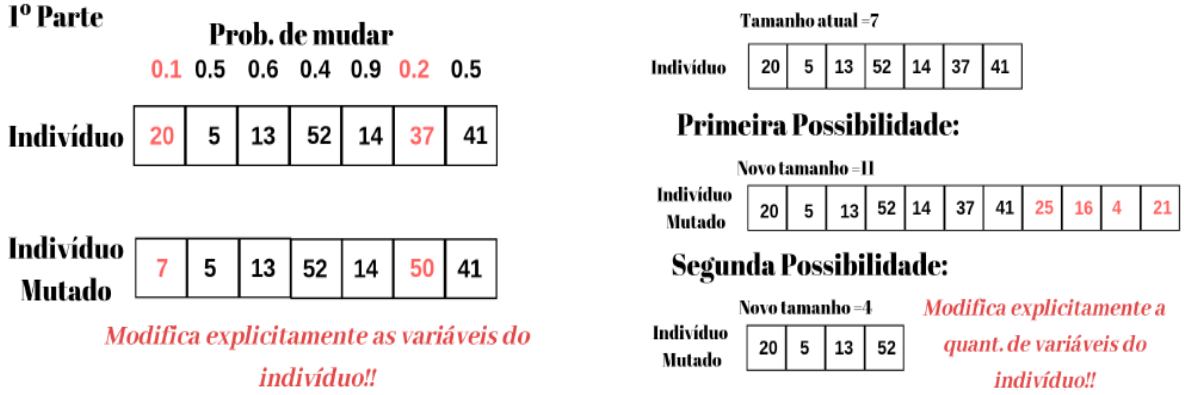



Figura 2 – Primeira e Segunda parte da mutação




Os valores das probabilidades dos operadores genéticos e a quantidade de indivíduos foram definidos após a realização de testes com vários valores e várias combinações entre eles. O número de gerações utilizado foi 50.

Este algoritmo busca minimizar os dois objetivos f1 e f2. O objetivo f1 se refere à quantidade de variáveis de uma solução. O segundo objetivo refere-se à taxa de erro de classificação dos dados utilizando o classificador KNN. O objetivo f1 de acordo a representação utilizada é calculado da seguinte forma:

$$f1 = |I| \quad \text{(Eq. 1)}$$

$$f2 = \frac{ErrFalhas}{ErrFalhas + AccFalhas} \quad \text{(Eq. 2)}$$

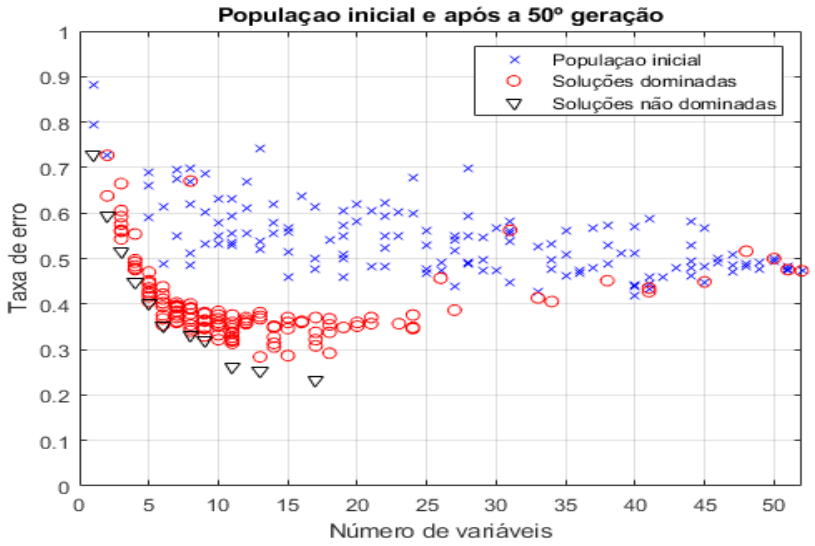
Onde *ErrFalhas* representa a quantidade de erros de classificação de falhas e *AccFalhas* representa a quantidade de acertos de classificação de falhas. O método escolhido para classificar as falhas foi o KNN (K-nearest neighbors). A escolha do KNN como método de classificação foi arbitrária, uma vez que para a metodologia proposta neste trabalho, qualquer método baseado em dados seria suficiente. O KNN possui uma etapa de predição onde ele utiliza os dados históricos como treinamento para atribuir a classe aos dados a serem classificados. Ele calcula a distância euclidiana da instância a ser classificada para todas as instâncias de treinamento e atribui a ela a classe da maioria entre as K com as primeiras menores distâncias. O valor K é um parâmetro do algoritmo e o valor utilizado foi 3, depois de testes com vários valores.

## Resultados Experimentais e Discussões

A performance de uma solução medida pela taxa de erro na classificação pelo KNN no conjunto de dados de teste do TEP se tornou inviável usando todos os dados disponíveis do TEP devido ao grande tempo demandado. Sendo assim, para treinamento utilizamos todos os 480 dados disponíveis para todos os 17 tipos de falhas testados. Porém, para os dados de teste o conjunto foi reduzido por meio de uma amostragem aleatória de 50 dados para cada tipo de falha. Do total de 21 falhas as 3,9,15 e 21 ,como explicado na subseção 2.1, não foram utilizadas. O algoritmo foi executado 30 vezes e a cada execução foi feita amostragens diferentes dos dados de teste. Embora a cada execução o algoritmo retorne uma população de indivíduos, foi selecionada a solução com a menor taxa de erro de classificação. A figura 3 abaixo ilustra, como exemplo, a população inicial e as soluções

dominadas e não-dominadas da população evoluída após a 50ª geração de uma execução.

Figura 3 – População inicial e população evoluída após 50ª geração.




Essas soluções foram analisadas e comparadas com as soluções selecionadas da mesma forma obtidas pelo algoritmo proposto por Silva *et al.* [2017], o NSGA-II-GMM-AP, que dentre os algoritmos analisados por eles, teve a menor taxa de erro.

No NSGA-II-GMM-AP é utilizada uma variação do *Non-domineted Sorting Gennetic Algorithm II* (NSGA-II) e classificadores baseado em Modelos de Mistura Gaussiana (GMM) e o conceito de Ponto de Atração para controlar a complexidade das soluções. A representação utilizada é binária, com uma população de 400 indivíduos. Para o cruzamento foram utilizados 2 pontos de cortes com probabilidade 0.9, e a mutação de um bit com probabilidade 0.1. A população foi evoluída durante 50 gerações e o algoritmo foi testado durante 30 execuções, e para cada execução foi feita uma amostragem de 50 instâncias dos dados de teste [Silva *et al.*, 2017].

A tabela 1 abaixo traz informações sobre as médias e os erros padrões (EP) do número de variáveis e do erro de classificação dos algoritmos.

Tabela 1 – Estatísticas das 30 execuções dos algoritmos.

| **Algoritmo** | **Média de Variáveis EP(%)** | **Média do Erro(%)EP(%)** |
|---|---|---|
| MOEADD-KNN-M | 12.903.79 | 27.020.86 |
| NSGA-II-GMM-AP | 32.470.55 | 18.830.18 |
| KNN | 52.000.00 | 52.450.52 |
| GMM | 52.000.00 | 23.170.19 |

O NSGA-II-GMM-AP levou a soluções com menores taxas de erro que o método proposto, embora o método proposto diminua o número de variáveis. Porém, quando se compara com a taxa de erro obtida pelo classificador KNN sem a seleção de variáveis (indicado na tabela 1 por KNN) percebe-se que a diminuição da taxa de erro para o método de seleção proposto foi significativa. Essa diminuição foi de 48,48%, maior do que a que houve entre o NSGA-II-GMM-AP em relação à taxa de erro dos classificadores GMM sem aplicação de seleção de variáveis (indicado na tabela 1 por GMM) que foi de 18,73%. As médias e os erros padrões das taxas de erro e número de variáveis do NSGAII- GMM-AP e do GMM sem seleção de variáveis foram retiradas de Silva *et al.* [2017].

## Conclusão

Este trabalho propôs um método wrapper de seleção de variáveis baseado no algoritmo evolutivo multiobjetivo MOEADD e no classificador KNN com operadores genéticos adaptados ao problema, denominado MOEADD-KNN-M, aplicados a classificação de falhas no TEP. O algoritmo proposto foi comparado ao algoritmo NSGA-II-GMM-AP e embora a taxa de erro das soluções do MOEADD-KNN-M tenha sido maiores, a proporção de diminuição da taxa de erro comparado com seu respectivo método de classificação sem a seleção de variáveis foi maior. Trabalhos futuros podem explorar o uso de outro classificador para o método wrapper proposto, como o GMM que se mostrou com uma taxa de erro menor do que o KNN sem a seleção de variáveis. O MOEADD pode ser explorado para uma versão tri-objetivo usando

outras métricas de performance do classificador, já que ele possui bom desempenho em mais de dois objetivos. O MOEADD-KNN-M pode ser aplicado a outros conjuntos de dados, ainda, com características distintas de sensibilidade a seleção de variáveis.

**Agradecimentos**

Agradecemos à Pró-Reitoria de Pesquisa da UNIMONTES, por nos proporcionar a realização desta pesquisa e a obtenção dos resultados. Agradecemos à UFMG-ICA por tornar pública esta pesquisa, em parceria com pesquisadores da UNIMONTES e UFMG-ICA.

**Financiamentos**

Todos os resultados obtidos na pesquisa foram obtidos com recursos próprios dos pesquisadores.

**Aprovação do Comitê de Ética**

## Referências

Ahmad, I. (2015). Feature selection using particle swarm optimization in intrusion detection.

International Journal of Distributed Sensor Networks, 11(10):806954.

Al-Ani, A., Alsukker, A., e Khushaba, R. N. (2013). Feature subset selection using differential

evolution and a wheel based search strategy. Swarm and Evolutionary Computation, 9:15–26.

Allegrini, F. e Olivieri, A. C. (2011). A new and efficient variable selection algorithm based on ant colony optimization. applications to near infrared spectroscopy/partial leastsquares analysis. Analytica chimica acta, 699(1):18–25.

Brusco, M. J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. Computational Statistics & Data Analysis, 77:38–53.

Chandrashekar, G. e Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1):16–28.

Foster, D., Karloff, H., e Thaler, J. (2015). Variable selection is hard. In Conference on Learning Theory, p. 696–709.

Guyon, I. e Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar):1157–1182.

Kohavi, R. e John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2):273–324.

Li, K., Deb, K., Zhang, Q., e Kwong, S. (2014). An evolutionary many-objective optimization algorithm based on dominance and decomposition. IEEE Transactions on Evolutionary Computation, 19(5):694–716.

Marill, T. e Green, D. (1963). On the effectiveness of receptors in recognition systems. IEEE transactions on Information Theory, 9(1):11–17.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., e Coello, C. A. C. (2013). A survey of multiobjective evolutionary algorithms for data mining: Part i. IEEE Transactions on Evolutionary Computation, 18(1):4–19.

Silva, F. M. S., Ferreira, J. F., Palhares, R. M., e D'Angelo, M. F. S. V. (2017). Uma abordagem evolutiva multiobjetivo baseada em ponto de atração para seleção de variáveis em problemas de classificação de falhas. XLIX Simpósio Brasileiro de Pesquisa Operacional.

Venkatasubramanian, V., Rengaswamy, R., e Kavuri, S. N. (2003a). A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies. Computers & chemical engineering, 27(3):313–326.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., e Yin, K. (2003b). A review of process fault detection and diagnosis: Part iii: Process history based methods. Computers & chemical engineering, 27(3):327–346.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., e Kavuri, S. N. (2003c). A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. Computers & chemical engineering, 27(3):293–311.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. IEEETransactions on Computers, 100(9):1100–1103.

Yu, L. e Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy.

Journal of machine learning research, 5(Oct):1205–1224.