11º FEPEG FÓRUM ENSINO · PESQUISA EXTENSÃO · GESTÃO
UNIVERSIDADE, SOCIEDADE E POLÍTICAS PÚBLICAS
ISSN: 1806-549X
Realização: Unimontes Universidade Estadual de Montes Claros; Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Ensino Superior; Minas Gerais Governo de Todos; PIBID Unimontes Programa Institucional de Bolsa de Iniciação à Docência
Apoio: FAPEMIG; FADENOR

UMA ABORDAGEM EVOLUTIVA MULTIOBJETIVO BASEADA EM PONTO DE ATRAÇÃO PARA SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CLASSIFICAÇÃO DE FALHAS

Autores: JESSICA FLAVIANE FERREIRA, FERNANDO MARCOS SOUZA SILVA, REINALDO MARTINEZ PALHARES, MARCOS FLÁVIO SILVEIRA VASCONCELOS D'ANGELO, TARCISIO FRANCISCO BATISTA FILHO

Introdução

Nas últimas décadas, o uso de sistemas computacionais trouxe grandes melhorias para o monitoramento de eventos anormais em processos, embora ainda haja parte significativa destas tarefas feitas por operadores humanos. A complexidade dos processos e o grande número de dados que devem ser analisados para uma tomada de decisão adequada são características impeditivas para a atuação integral de humanos nestas tarefas, ao passo que se tornam um convite para a automação das atividades desta área denominada Gerenciamento de Eventos Anormais, do inglês Abnormal Event Management (AEM) (VENKATASUBRAMANIAN et al., 2003a).

Um dos componentes das atividades de AEM são os Sistemas de Detecção e Diagnóstico de Falhas, do inglês Fault Detection and Diagnosis (FDD), que a partir do monitoramento de um processo, apontam possíveis falhas e as diagnosticam, classificando-as. (VENKATASUBRAMANIAN et al., 2003a,b,c). Classificação de falhas consiste em discriminar os tipos das falhas detectadas pelo sistema, i.e., distinguir as causas de uma condição anormal. Esta é uma tarefa importante, pois a partir da correta identificação de uma falha em tempo hábil é possível dar início a atividades corretivas evitando o comprometimento da produtividade e segurança de um processo. Uma falha não detectada e/ou não classificada em tempo hábil pode ter efeitos catastróficos.

Os processos industriais modernos estão se tornando cada vez mais complexos, tanto em níveis estruturais quanto de automação, para aplicações de FDD, podendo ter muitas variáveis redundantes e/ou irrelevantes lidas diretamente de sensores dispersos pelas plantas. A eliminação destas variáveis do conjunto de variáveis monitoradas pode levar a uma redução de custos de monitoramento e também a sistemas de FDD mais eficientes e robustos. Entretanto, a escolha do conjunto de variáveis mais adequado para o monitoramento não é uma tarefa fácil (FOSTER et al., 2014).

O problema aqui apresentado, abordado na literatura na área de mineração de dados como seleção de variáveis, por ser NP-hard de complexidade $O(2n)$, tem sido frequentemente tratado por meio de métodos heurísticos ou metaheurísticos, que são largamente classificados como filter e wrapper (CHANDRASHEKAR; SAHIN, 2014). As estratégias mais utilizadas podem ser mono-objetivo (quando busca-se minimizar o erro de classificação) ou multiobjetivo, que normalmente buscam minimizar o erro de classificação ao mesmo tempo que minimizam a complexidade das soluções encontradas, i.e., o número de variáveis selecionadas.

Neste trabalho é feita uma proposta de um método wrapper multiobjetivo de seleção de variáveis para aplicação em classificação de falhas baseada em dados históricos de um processo petroquímico usando um algoritmo genético multiobjetivo e classificadores baseados em modelos de Mistura Gaussiana, do inglês Gaussian Mixture Models (GMM). No método proposto, denominado NSGA-II-GMM-AP, é utilizada uma variação do Non-dominated Sorting Genetic Algorithm II (NSGA-II) (DEB et al., 2002) como mecanismo de busca.

A principal característica do método proposto é que ele utiliza o conceito de Ponto de Atração visando controlar a complexidade das soluções da população mantida pelo NSGA-II durante o processo de otimização, mantendo-a próxima à da melhor solução obtida em cada geração. Isto é feito buscando tratar o problema de que os operadores de cruzamento tradicionais em algoritmos evolutivos tendem a explorar soluções de complexidade média em problemas de seleção de variáveis (EMMANOUILIDIS et al., 2000). Outra característica do método é que apesar de seguir uma estratégia multiobjetivo de busca por ser baseado no NSGA-II, o NSGA-II-GMM-AP retorna apenas uma única solução de melhor acurácia, por que nós argumentamos que este é o objetivo mais importante a ser otimizado em um sistema crítico de FDD.

Apoio Financeiro: CNPq e FAPEMIG.

ISSN: 1806-549X

Material e métodos

A. Classificadores GMM e o Processo Tennessee Eastman

Os classificadores baseados em GMM são métodos de classificação supervisionada de falhas compostos por uma etapa de treinamento (offline), em que os parâmetros do GMM são otimizados para cada tipo de falha do processo; e uma etapa de predição (online), em que por meio do valor da função de densidade de probabilidade, do inglês Probability Density Function (PDF), calculada a probabilidade de um novo dado pertencer a cada tipo de falha. A escolha da classe de falha é feita baseando-se no maior valor de PDF para esse dado.

O Tennessee Eastman Process (TEP), no qual os métodos deste trabalho foram aplicados, é um processo industrial petroquímico muito utilizado para avaliar métodos de controle de processos e monitoramento. O TEP é descrito no trabalho de Downs e Vogel (1993) e revisado por Bathelt et al. (2015). Neste trabalho foram utilizados os dados do TEP disponíveis no site do Massachusetts Institute of Technology (MIT) que contém 52 variáveis e 21 tipos de falha. Partindo-se do pressuposto de que as falhas já foram detectadas pelo sistema de FDD, tratar-se-á neste trabalho o problema de classificação de falhas usando GMM. Foram descartados os dados de condições de operação normal, do inglês Normal Operation Condition (NOC), e utilizadas 480 instâncias para treinamento e 800 para teste de cada tipo de falha.

B. NSGA-II-GMM-AP

Foram implementados neste trabalho 5 métodos wrapper de seleção de variáveis, todos com GMM: o método NSGA-II-GMM, que é o NSGA-II aplicado ao problema de seleção de variáveis usando classificadores GMM; o método NSGA-II-GMM-AP, que é o NSGA-II-GMM citado utilizando o conceito de Ponto de atração, do inglês Attraction Point (AP), para busca do controle de complexidade das soluções; o MONO-GA-GMM, que é um algoritmo genético mono-objetivo com GMM; o SFS-GMM, que é o método Sequential Forward Selection (SFS) com GMM; e o SBS-GMM, que é o método Sequential Backward Selection (SBS) com GMM. Os três últimos métodos foram implementados para fins de comparação.

O NSGA-II-GMM possui indivíduos com representação binária, sendo cada um composto por um vetor de tamanho 52 (quantidade de variáveis do TEP). Cada gene do indivíduo possui a informação sobre a seleção (valor um) ou não seleção (valor zero) da variável. A população é composta por 400 indivíduos, inicializados de forma aleatória, seguindo uma distribuição uniforme da quantidade de variáveis que compunham os indivíduos. Neste algoritmo foi utilizado o operador de cruzamento com dois pontos de corte, probabilidade de cruzamento de 90% e o método de seleção foi o de torneio binário. O critério de parada é de 50 gerações.

Um problema recorrente que ocorre em algoritmos evolutivos quando se utiliza operadores tradicionais de cruzamento que levam à exploração de soluções de complexidade média é descrito por Emmanouilidis et al. (2000). Solução de complexidade média quer dizer que o cruzamento de uma solução de n variáveis com outra solução de m variáveis tende a gerar uma solução com $(m+n)/2$ variáveis. Assim, foi proposto neste trabalho um novo método de cruzamento controlando-se a complexidade das soluções por meio de ponto de atração, ao se manter as soluções com complexidades próximas das que possuem a melhor performance. Com o objetivo de sanar estes problemas, propôs-se o NSGA-II-GMM-AP. Este algoritmo foi desenvolvido com as mesmas características do NSGA-II-GMM com algumas modificações. A mais evidente é com relação ao objetivo de minimização $f1$, que representa a distância entre a complexidade do indivíduo I que possui n genes e a complexidade do indivíduo com menor erro de classificação presente na população. As funções $f1$ e $f2$ na Fig. 1 representam os objetivos a serem minimizados, em que I é a complexidade do indivíduo com n genes, AP é a complexidade do indivíduo com menor erro de classificação, FP são os falsos positivos, FN os falsos negativos, TP são os verdadeiros positivos e TN são os verdadeiros negativos.

Resultados e discussão

Em todos os algoritmos desenvolvidos a performance dos classificadores GMM é medida pelo erro de classificação no conjunto de dados de testes do TEP. Os critérios utilizado para comparação entre os métodos implementados foram o erro de classificação e a quantidade de variáveis das soluções. Todos os algoritmos foram executados 30 vezes, portanto, foram feitas 30 amostragens diferentes dos dados de testes. Ao término de cada execução de cada algoritmo, foi selecionada a solução com menor erro de classificação, por considerarmos que este seja o principal critério para seleção de variáveis em sistemas críticos de FDD. Destas soluções escolhidas, analisou-se o erro de classificação e a quantidade de variáveis selecionadas. A Tabela 1 traz informações sobre as médias e os erros padrões (EP) do número

Apoio Financeiro: CNPq e FAPEMIG

ISSN: 1806-549X

Ao comparar a performance dos métodos de seleção de variáveis com os classificadores GMM quando não se faz seleção de variáveis é possível perceber o benefício que a seleção de variáveis pode trazer para os sistemas de FDD, graças à diminuição do erro de classificação e do número de variáveis selecionadas pelos algoritmos.

Conclusões

Este trabalho teve como objetivo a proposta e estudo de um método wrapper evolutivo multiobjetivo de seleção de variáveis, baseado no algoritmo NSGA-II, classificadores usando GMM e o conceito de ponto de atração, denominado NSGA-II-GMM-AP, aplicado ao problema de classificação de falhas no TEP. Esse método foi comparado a outros métodos de seleção de variáveis. Os resultados obtidos mostraram que o NSGA-II-GMM-AP leva a soluções com menores erros de classificação que os algoritmos usados para comparação no conjunto de dados testado.

Agradecimentos

Ao CNPq e à FAPEMIG pelo apoio financeiro por meio do PIBIC.

Referências bibliográficas

BATHELT, A.; RICKER, N. L.; JELALI, M. Revision of the tennessee eastman process model. **IFAC-PapersOnLine**, v. 48, n. 8, p. 309–314, 2015.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, n. 1, p. 16–28, 2014.

DEB, K.; PRATAP, A.; AGARWAL, S.; MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE transactions on evolutionary computation** , v. 6, n. 2, p. 182–197, 2002.

DOWNS, J. J.; VOGEL, E. F. A plant-wide industrial process control problem. **Computers & chemical engineering**, v. 17, n. 3, p. 245–255, 1993.

EMMANOUILIDIS,C.; HUNTER, A.; MACINTYRE, J. A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. In **Evolutionary Computation**, 2000. Proceedings of the 2000 Congress on, v. 1, p. 309–316. IEEE.

FOSTER, D. P.; KARLOFF, H. J.; THALER, J. **Variable selection is hard**. 2014. CoRR, abs/1412.4832. Disponível em: <https://arxiv.org/abs/1412.4832>. Acesso em: 27 set. 2017

SILVA, F. M. S.; FERREIRA, J. F.; PALHARES, R. M.; D'ANGELO, M. F. S. V. **Uma abordagem evolutiva multiobjetivo baseada em ponto de atração para seleção de variáveis em problemas de classificação de falhas.** Pré-anais do XLIX Simpósio de Brasileiro de Pesquisa Operacional, 2017. Disponível em: . Acesso em: 27 set. 2017.

VENKATASUBRAMANIAN, V.; RENGASWAMY, R.; YIN, K.; KAVURI, S. N. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. **Computers & chemical engineering,** v. 27, n. 3, p. 293-311, 2003a.

VENKATASUBRAMANIAN, V.; RENGASWAMY, R.; KAVURI, S. N. A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies. **Computers & Chemical Engineering,** v. 27, n. 3, p. 313-326, 2003b.

VENKATASUBRAMANIAN, RENGASWAMY, R.; KAVURI, S. N.; YIN, K. A review of process fault detection and diagnosis: Part iii: Process history based methods. **Computers & chemical engineering,** v. 27, n. 3, p. 327-346, 2003c.

Tabela 1. Estatísticas das 30 execuções dos algoritmos.

Algoritmo	Média das variáveis ± EP(%)	Média de Erro ± EP(%)
NSGA-II-GMM	23,17 ± 0,70	19,89± 0,17
NSGA-II-GMM-AP	32,47 ± 0,55	18,83 ± 0,18
MONO-GA-GMM	35,17 ± 0,70	19,27 ± 0,17

Apoio Financeiro: CNPq e FAPEMIG.

ISSN: 1806-549X

Algoritmo	Média das variáveis ± EP(%)	Média de Erro ± EP(%)
SFS-GMM	33,30 ± 1,32	20,27 ± 0,26
SBS-GMM	26,97 ± 0,89	19,67 ± 0,18
GMM	52,00 ± 0,00	23,17 ± 0,19

$$f1: \left| \sum_{i=1}^{n} I_i - AP \right|$$

$$f2: \frac{FP+FN}{TP+FP+TN+FN}$$

Figura 1: Funções-objetivo a serem minimizadas.

Apoio Financeiro: CNPq e FAPEMIG.