UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA INDUSTRIAL - PEI

MESTRADO EM ENGENHARIA INDUSTRIAL

KARINE DO PRADO RIBEIRO

CONTRIBUIÇÕES PARA A APLICAÇÃO DE ALGORITMO GENÉTICO NO AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS

PEI
Programa de Pós-Graduação em Engenharia Industrial

SALVADOR
2018

UFBA

UNIVERSIDADE FEDERAL DA BAHIA

ESCOLA POLITÉCNICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA INDUSTRIAL – PEI

MESTRADO EM ENGENHARIA INDUSTRIAL

KARINE DO PRADO RIBEIRO

CONTRIBUIÇÕES PARA A APLICAÇÃO DE ALGORITMO GENÉTICO NO AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS

Salvador, 2018

KARINE DO PRADO RIBEIRO

CONTRIBUIÇÕES PARA A APLICAÇÃO DE ALGORITMO GENÉTICO NO AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Industrial, da Universidade Federal da Bahia, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Industrial.

Orientador: Prof. Dr. Cristiano Fontes

Salvador
2018

Ribeiro, Karine do Prado

Contribuições para a Aplicação de Algoritmo Genético no Agrupamento e Classificação de Séries Temporais Multivariadas / Karine do Prado Ribeiro. -- Salvador, 2018.

107 f. : il

Orientador: Cristiano Hora Fontes.

Dissertação (Mestrado – Programa de Pós-Graduação em Engenharia Industrial – PEI) -- Universidade Federal da Bahia, Escola Politécnica, 2018.

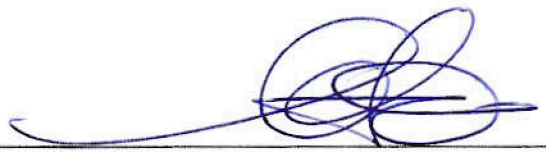
1. Algoritmo Genético. 2. Agrupamentos. 3. Séries Temporais Multivariadas. 4. Reconhecimento de Padrões. 5. Detecção de Falhas. I. Fontes, Cristiano Hora. II. Título.

Contribuições para a Aplicação de Algoritmo Genético no Agrupamento e Classificação de Séries Temporais Multivariadas

Karine do Prado Ribeiro

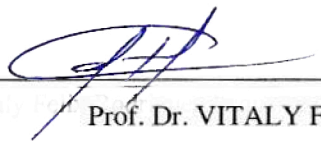
Dissertação submetida ao corpo docente do programa de pós-graduação em Engenharia Industrial da Universidade Federal da Bahia como parte dos requisitos necessários para a obtenção do grau de mestre em Engenharia Industrial.

Examinada por:



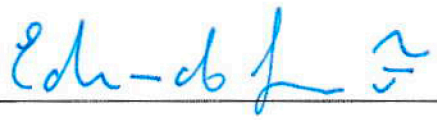
Prof. Dr. CRISTIANO HORA FONTES (Orientador)

Doutor em Engenharia Química pela Universidade Estadual de Campinas, UNICAMP, Brasil, 2001



Prof. Dr. VITALY FÉLIX RODRÍGUEZ ESQUERRE

Doutor em Engenharia Elétrica pela Universidade Estadual de Campinas, UNICAMP, Brasil, 2003



Prof. Dr. EDUARDO FURTADO DE SIMAS FILHO

Doutor em Engenharia Elétrica pela Universidade Federal do Rio de Janeiro, UFRJ, Brasil, 2010

Salvador, BA - BRASIL

Dezembro/2018

Dedico este trabalho a Deus, família, amigos e ao meu orientador pelo apoio, força, incentivo, companheirismo e amizade. Sem eles nada seria possível.

Agradecimentos

- A Deus por ter me dado saúde e força interior para superar as dificuldades.

- Ao meu orientador, professor Cristiano Hora Fontes, por todo empenho, sabedoria e compreensão.

- Ao meu noivo e companheiro, Gabriel Jesus Alves de Melo, por todos os dias que esteve ao meu lado, pela cumplicidade e afeto, pelo incentivo, por me escutar e me ajudar sempre.

- As minhas amigas, Aline Ramos, Ana Célia Barreto e Nina Santiago, por suas companhias, carinho, incentivo, conselhos e bons momentos de descontração.

- As funcionárias da secretaria do PEI, Tatiane e Tamiles, pela eficiência, atenção, cuidado e carinho que sempre tiveram comigo.

- Aos meus pais, Manoel Milton Ribeiro e Gleide do Prado Ribeiro e ao meu irmão, Leandro do Prado Ribeiro, pelo amor e carinho incondicional e pelos ensinamentos obtidos, estes estarão comigo por toda vida.

- Aos professores, funcionários e colegas do Programa de Pós-graduação em Engenharia Industrial da UFBA.

"A persistência é o caminho do êxito." (Charles Chaplin).

Resumo da Dissertação apresentada ao PEI/UFBA como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Industrial (MSEng.)

CONTRIBUIÇÕES PARA A APLICAÇÃO DE ALGORITMO GENÉTICO NO AGRUPAMENTO E CLASSIFICAÇÃO DE SÉRIES TEMPORAIS MULTIVARIADAS

Karine do Prado Ribeiro

Dezembro/2018

Orientador: Prof. Dr. Cristiano Hora de Oliveira Fontes.

A análise de agrupamentos e o reconhecimento de padrões a partir de dados de processo representa uma alternativa para a extração de conhecimento útil e, entre outros, para a detecção e diagnóstico de falhas (FDD, *Fault Detection and Diagnosis*). De forma inovadora, este trabalho apresenta uma metodologia voltada ao reconhecimento de padrões em séries temporais multivariadas que consiste na adaptação e aplicação de Algoritmos Genéticos (GA, *Genetic Algorithm*) em um método clássico de agrupamento não hierárquico baseado em otimização (FCM, *fuzzy c-means*). A análise de similaridade entre as séries, coletadas em diferentes períodos de operação (doravante aqui denominados de objetos), é realizada com base em duas métricas das quais uma compara a direção dos componentes principais entre os objetos (SPCA, ou PCA *Similarity factor*) e a outra compara os seus respectivos valores médios ou níveis de operação (AED, *Average-based Euclidean Distance*). Dois estudos de caso foram analisados, ambos inspirados em problemas de detecção de falhas em processos de produção. A primeira aplicação compreendeu um processo industrial real relacionado à operação de partida de uma turbina a gás de escala comercial (Unidade Termoelétrica Rômulo Almeida, Petrobras). O segundo estudo de caso envolveu o reconhecimento de padrões em eventos normais e de falha a partir de séries multivariadas extraídas de uma unidade virtual de referência utilizada na análise de estratégias de controle e FDD (*Tennessee Eastman Process* – TEP). Este trabalho evidencia as potencialidades de aplicação de um método heurístico de otimização em relação à abordagem de otimização clássica para a resolução de um problema de agrupamento envolvendo séries

multivariadas. O melhor desempenho da estratégia heurística (GA) se verifica principalmente através da melhor exploração da região de busca e da obtenção de um melhor mínimo local do ponto de vista da qualidade da classificação. Os resultados obtidos mostram que o FCM baseado em GA apresentou um percentual de acerto de classificação igual ou superior ao método FCM baseado em otimização clássica, o que comprova a viabilidade da alternativa proposta para a codificação dos genes e a eficácia da abordagem heurística em problemas que sugerem a existência de múltiplos mínimos locais. A estratégia proposta para a aplicação de algoritmos genéticos no agrupamento e reconhecimento de padrões em séries multivariadas representa uma alternativa potencial para a extração de conhecimento de um processo de produção, para o apoio à tomada de decisão na gestão operacional e para implementação de estratégias de controle ótimo através dos padrões reconhecidos.

Abstract of Dissertation presented to PEI/UFBA as a partial fulfillment of the requirements for the degree of Master of Industrial Engineering (MSEng.)

CONTRIBUTIONS TO THE APPLICATION OF A GENETIC ALGORITHM IN THE CLUSTERING AND CLASSIFICATION OF MULTIVARIATE TIME SERIES

Karine do Prado Ribeiro

Dezembro/2018

Advisor: Prof. Dr. Cristiano Hora de Oliveira Fontes.

Cluster analysis and pattern recognition from process data represent an alternative for the extraction of useful knowledge and, among others, for the detection and diagnosis of faults (FDD, Fault Detection and Diagnosis). In an innovative way, this study presents a methodology for the recognition of patterns in multivariate time series, which consists of the customization and application of Genetic Algorithms (GA) in a classical non-hierarchical clustering method based on optimization (FCM, fuzzy c-means). The similarity analysis between the time series, collected in different periods of operation (objects), is performed based on two metrics of which one compares the direction of the main components between the objects (SPCA, or PCA Similarity factor) and the other compares their respective average values or levels of operation (AED, Average-based Euclidean Distance). Two case studies were analyzed, both inspired by problems of detection of faults in production processes. The first application comprised an actual industrial process related to the start-up operation of a commercial scale gas turbine (Thermoelectric Unit, Petrobras). The second case study involved the recognition of patterns in both normal and fault events from multivariate series extracted from a virtual reference unit used in the analysis of control strategies and FDD (Tennessee Eastman Process - TEP). This study highlights the potentialities of applying a heuristic optimization method in relation to the classical optimization approach to solve a clustering problem involving multivariate series. The best performance of the heuristic strategy (GA) is verified mainly by expanding the search region and obtaining a better local minimum from the point of view of the classification quality. The results show that the FMC based on GA presented a percentage of classification accuracy equal

to or greater than the FCM method based on classical optimization, which proves the viability of the proposed alternative for the coding of the genes and the effectiveness of the heuristic approach in problems that suggest the existence of multiple local minimums. The proposed strategy for the application of genetic algorithms in clustering and pattern recognition in multivariate series represents a potential alternative for the extraction of knowledge from a production process as well as support for decision making in operational management and for implementation of great control strategies using recognized patterns.

Lista de Figuras

Figura 1 – Um esquema do loop de monitoramento de processo (CHIANG, et al., 2000). *28*
Figura 2 – Agrupamento 20 cidades em dois grupos. *30*
Figura 3 – Agrupamentos rígido e fuzzy (Adaptado de JAIN et al., 1999). *35*
Figura 4 – Pessoas alfabetizadas de 5 anos ou mais de idade no Brasil. *38*
Figura 5 – Representação de séries temporais multivariadas. *39*
Figura 6 – Um algoritmo genético simples (HOUCK et al., 1995). *44*
Figura 7 – Representação de um cromossomo. *52*
Figura 8 – Operações em cada iteração do GA (RIBEIRO e FONTES, 2017). *53*
Figura 9 – Técnica do tipo roleta para seleção de cromossomos. *56*
Figura 10 – Roleta com 20 posições. *57*
Figura 11 – Crossover. *59*
Figura 12 – Mutação. *60*
Figura 13 – Termelétrica Rômulo Almeida (SÁ BARRETTO, 2009). *62*
Figura 14 – Turbina a gás RB211-G62 (ROLLS-ROYCE, 2010). *63*
Figura 15 – Etapas da metodologia – caso 1. ..... *65*
Figura 16 – Séries da amostra de treinamento. *67*
Figura 17 – Exemplo de partida normal. *68*
Figura 18 – Exemplo de partida com falha. *68*
Figura 19 – Estabilização da aptidão da melhor solução ao longo das gerações. *69*
Figura 20 – Centro com oscilações excessivas (ausência de limitação para as alterações de genes através de mutação). *70*
Figura 21 – Distribuição dos objetos nos três grupos. *71*
Figura 22 – Padrão normal 1 obtido (FCM-GA). *72*
Figura 23 – Padrão normal 2 (FCM-GA). *73*
Figura 24 – Padrão de falha (FCM-GA). *73*
Figura 25 – Soma das distâncias extragrupos (FCM-GA). *74*
Figura 26 – Soma das distâncias intragrupos (FCM-GA). *74*
Figura 27 – Padrão normal obtido com o FCM clássico. *76*
Figura 28 – Padrão de falha obtido com o FCM clássico. *76*
Figura 29 – Função objetivo (FCM com otimização clássica) ao longo das iterações. *77*
Figura 30 – Distâncias entre os padrões reconhecidos. *77*
Figura 31 – TEP (DOWNS, VOGEL, 1993). *79*
Figura 32 – Sistema de controle aplicado ao TEP (RICKER, 1996). *81*
Figura 33 – Séries da amostra de treinamento – TEP falha 2. ..... *84*
Figura 34 – Objetos distribuídos nos três grupos – TEP falha 2. ..... *85*
Figura 35 – Padrão normal 1 obtido (FCM-GA) – TEP falha 2. ..... *86*
Figura 36 – Padrão normal 2 obtido (FCM-GA) – TEP falha 2. ..... *86*
Figura 37 – Padrão de falha obtido (FCM-GA) – TEP falha 2. ..... *87*
Figura 38 – Padrão normal 1 obtido com o FCM clássico – TEP falha 2. ..... *88*

Figura 39 – Padrão normal 2 obtido com o FCM clássico – TEP falha 2. ..... *88*
Figura 40 – Padrão de falha obtido com o FCM clássico – TEP falha 2. ..... *89*
Figura 41 – Distâncias entre os padrões reconhecidos – TEP falha 2. ..... *90*
Figura 42 – Algumas séries da amostra (variável: temperatura). *91*
Figura 43 – Algumas séries da amostra (variável: vazão de líquido). *92*
Figura 44 – Padrão normal obtido com o GA – TEP falha 3. ..... *93*
Figura 45 – Padrão de falha obtido com o GA – TEP falha 3. ..... *93*
Figura 46 – Padrão normal obtido com o FCM otimização clássica – TEP falha 3. ..... *94*
Figura 47 – Padrão de falha obtido com o FCM otimização clássica – TEP falha 3. ..... *95*

Lista de Tabelas

Tabela 1 – Resumo dos trabalhos de agrupamento de séries temporais baseados em GA *49*
Tabela 2 – Valores dos parâmetros do algoritmo genético *60*
Tabela 3 – Amostras de treinamento e de teste *70*
Tabela 4 – Porcentagem de classificações erradas (FCM-GA) *72*
Tabela 5 – Porcentagem de classificações erradas – FCM com otimização clássica *75*
Tabela 6 – Porcentagem de classificações erradas (TEP falha 2) (FCM-GA) *85*
Tabela 7 – Porcentagem de classificações erradas (TEP falha 2) – otimização clássica *87*
Tabela 8 – Porcentagem de classificações erradas (TEP falha 2 – alteração na estimativa inicial) – FCM-GA *90*
Tabela 9 – Porcentagem de classificações erradas (TEP falha 2 – alteração na estimativa inicial) – otimização clássica *91*
Tabela 10 – Porcentagem de classificações erradas (TEP falha 3) – GA/FCM *94*

Lista de Abreviaturas

AED	*Average-based Euclidean Distance*
CA	Caldeira Auxiliar
FCM	*Fuzzy C-Means*
FDD	*Fault Detection and Diagnosis*
GA	*Genetic Algorithm*
HRSG	*Heat Recovery Steam Generator*
ONS	Operador Nacional do Sistema Elétrico
PCA	*Principal Component Analysis* (Análise de Componentes Principais)
PFD	*Process Fault Diagnosis*
PIMS	*Process Information Management System*
RNA	Rede Neural Artificial
SIN	Sistema Interligado Nacional
SPCA	*Principal Component Analysis Similarity Factor*
STM	Série Temporal Multivariada
STU	Série Temporal Univariada
TEP	*Tennessee Eastman Process*
TG	Turbina a Gás
TI	Tecnologia da Informação

TV	Turbina a Vapor
UFBA	Universidade Federal da Bahia
UTE-RA	Unidade Termelétrica Rômulo Almeida

Sumário

CAPÍTULO 1. INTRODUÇÃO 19

1.1 CONSIDERAÇÕES INICIAIS 19

1.2 OBJETIVOS 23

CAPÍTULO 2. FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA 24

2.1 RECONHECIMENTO DE PADRÕES 24

2.1.1 Visão geral 24

2.1.2 Detecção e diagnóstico de falhas (FDD) 27

2.2 ANÁLISE DE AGRUPAMENTOS 29

2.2.1 Visão geral 29

2.2.2 Principais etapas 31

2.2.3 Características 31

2.2.4 O agrupamento Fuzzy 33

2.2.4.1 Fuzzy c-means 35

2.3 SÉRIES TEMPORAIS 37

2.3.1 Agrupamento e Reconhecimento de Padrões em Séries Temporais 39

2.4 ALGORITMO GENÉTICO 43

2.4.1 Representação dos Cromossomos 44

2.4.2 População Inicial, Término e Função de Avaliação 45

2.4.3 Operadores Genéticos 45

2.4.3.1 Seleção 46

2.4.3.2 Crossover 47

2.4.3.3 Mutação 47

2.4.4 Agrupamento e Reconhecimento de Padrões com Algoritmo Genético 48

CAPÍTULO 3. METODOLOGIA 51

3.1 APLICAÇÃO DO ALGORITMO GENÉTICO PARA O AGRUPAMENTO E RECONHECIMENTO DE PADRÕES EM SÉRIES MULTIVARIADAS 51

CAPÍTULO 4. ESTUDOS DE CASO E RESULTADOS 62

4.1 ESTUDO DE CASO 1: DETECÇÃO DE FALHAS EM UMA TURBINA A GÁS . 62

4.1.1 Descrição 62

4.1.2 Procedimento e coleta de dados 65

4.1.3 Resultados e Discussão 66

4.2 ESTUDO DE CASO 2: DETECÇÃO DE FALHAS EM UMA UNIDADE VIRTUAL 78

4.2.1 Descrição 78

4.2.2 Coleta de Dados e Resultados 82

4.2.2.1 Falha 2 83

4.2.2.2 Falha 3 91

CAPÍTULO 5. CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS 96

5.1 CONCLUSÕES 96

5.2 SUGESTÕES PARA TRABALHOS FUTUROS 97

5.3 PUBLICAÇÃO 97

REFERÊNCIAS 98

CAPÍTULO 1
INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Os avanços verificados na tecnologia da informação (TI) permitiram a coleta e o armazenamento de grandes volumes de dados pelas organizações. Este cenário motivou o desenvolvimento de métodos associados à extração de conhecimento a partir de dados. A mineração de dados (*data mining*) é composta por um conjunto de técnicas que analisam grandes conjuntos de dados visando extrair conhecimento útil (KANTARDZIC, 2011). A partir da *data mining* é possível, por exemplo, descobrir regras, padrões e tendências de comportamento em processos e sistemas.

Técnicas de Mineração de Dados são amplamente utilizadas nas mais diversas áreas tais como *marketing* (MORO *et al*., 2011), geografia (MILLER, HAN, 2009), *business* (BERSON, SMITH, 2002), saúde (CHEN *et al*., 2017), inteligência artificial (ZHANG *et al*., 2017) e, sobretudo, em processos industriais (THOMAS *et al*., 2017; IZAKIAN *et al*., 2015; BANKÓ, ABONYI, 2012). Dentre estas técnicas se destaca o agrupamento de dados que viabiliza o reconhecimento de objetos similares (sem rótulos predefinidos) e o reconhecimento de padrões (ou protótipos) representativos para cada grupo ou classe identificada.

A análise de agrupamentos exerce um importante papel na interpretação e classificação de grande quantidade de objetos. O objetivo consiste na identificação de grupos (ou classes) cujos elementos internos sejam bastante similares entre si e dissimilares quando comparados aos elementos de outros grupos (OLIVEIRA *et al*., 2007).

A análise de agrupamentos pertence à categoria de problemas de aprendizado não supervisionado porque a informação do rótulo de classe de cada objeto não é considerada na formulação do problema. O agrupamento é uma forma de aprendizagem

por meio da observação e posterior extração de características, e não por meio de exemplos ou referências disponibilizadas a priori (HAN *et al*., 2011).

Dentre os métodos de agrupamento, destacam-se os métodos não hierárquicos de particionamento que requerem a definição prévia de um número de grupos (não rotulados) e a posterior classificação ou definição do grau de pertinência de cada objeto a cada um dos grupos. A partição é rígida (*crisp*) se cada objeto pertence apenas a um grupo e, neste caso, com grau de pertinência igual à unidade. A existência de sobreposição entre os grupos e a consequente incerteza na definição das fronteiras estabelece a necessidade de um particionamento *fuzzy* (nebuloso) no qual um objeto pode pertencer a mais de um grupo simultaneamente com diferentes graus de pertinência (LIAO, 2005; JAIN, 2010; GARAI, CHAUDHURI, 2004). Dentre os métodos de partição *fuzzy*, destaca-se o algoritmo *fuzzy c-means* (FCM) (BEZDEK, 2013) que compreende essencialmente um método de agrupamento baseado em otimização.

Os algoritmos de agrupamento não-hierárquicos, baseados em métodos clássicos de otimização (*k-means* e *fuzzy c-means*) possuem algumas desvantagens. A principal é que o processo de busca pela solução pode ser finalizado em um mínimo local que pode resultar em uma qualidade de agrupamento insatisfatória com uma elevada taxa de erro de classificação. Outra desvantagem é a inexistência de uma avaliação da qualidade dos grupos reconhecidos a cada iteração e a utilização desta informação nas iterações seguintes (RAHMAN, ISLAM, 2014). Tentando superar essas limitações, versões modificadas de algoritmos clássicos de agrupamento baseados em otimização têm sido propostas envolvendo o uso de outros algoritmos tais como Algoritmos Genéticos (GA) (BEZDEK, HATHAWAY, 1994; WHITLEY, 1994; LIU, XIE, 1995; MAULIK, BANDYOPADHYAY, 2000; MAULIK, BANDYOPADHYAY, 2003; ZHANG *et al*., 2007; SÁEZ *et al*., 2008; MUKHOPADHYAY *et al*., 2009) e Otimização por Enxame de Partículas (PSO) (LI *et al*., 2007; IZAKIAN *et al*., 2009; RUNKLER, KATZ, 2006; LI *et al*., 2012).

O GA tem se mostrado como um algoritmo eficaz de tentativa de otimização global devido a sua capacidade de explorar significativamente a região de busca (MENG *et al*., 2002). Os algoritmos genéticos são inspirados na teoria da evolução e visam ao longo de sucessivas iterações manter e aprimorar um conjunto de cromossomos (ou indivíduos) que formem uma população de possíveis soluções ótimas.

Uma função de avaliação examina cada cromossomo, de modo a verificar a qualidade de um indivíduo candidato à solução. As soluções são ordenadas através de uma função de avaliação por um critério de qualidade. Por meio desse critério, é possível selecionar os melhores indivíduos que terão maior probabilidade de propagar sua informação genética nas próximas iterações e, desta forma, gerar novas soluções ainda melhores que as atuais (BANDYOPADHYAY, MAULIK, 2002).

Dentre os diversos tipos de objetos que podem ser agrupados, as séries temporais são amplamente utilizadas considerando, entre outros, a sua capacidade de oferecer informação sobre a dinâmica dos processos (LIAO, 2005). A classificação ou agrupamento de séries temporais é capaz de fornecer informações úteis em diversas áreas, tais como processos industriais que armazenam seu conhecimento em extensas bases de dados históricas (ABONYI *et al*., 2005). O agrupamento de séries temporais pode ser baseado diretamente nos dados brutos, em características ou modelos empíricos identificados a partir das próprias séries (RANI, SIKKA, 2012).

Alguns fatores influenciam na escolha do algoritmo de agrupamento de séries temporais: período de amostragem uniforme ou não, séries univariadas ou multivariadas, comprimento uniforme ou não da janela de tempo das diferentes séries temporais (FU, 2011). Um dos desafios intrínsecos ao agrupamento de séries temporais, univariadas (STU) ou multivariadas (STM), compreende a escolha da métrica de similaridade. No caso univariado a distância Euclidiana é a métrica mais adotada (XUN, ZHISHU, 2010; WANG *et al*., 2013). Por outro lado, a distância Euclidiana não é geralmente recomendada para STM visto que cada objeto é tratado como uma simples coleção de séries univariadas (BANKÓ, ABONYI, 2012), não se efetuando uma análise do comportamento integrado dos perfis das variáveis contempladas em cada objeto. Dentre as métricas de similaridade aplicadas a séries multivariadas, tem-se o PCA (*Principal Component Analysis*) *similarity metric* (SPCA *index*), ou fator de similaridade baseado em PCA, o qual consiste em quantificar o grau de semelhança entre dois objetos (duas séries multivariadas) através da similaridade (ou dissimilaridade) entre as direções dos seus respectivos componentes principais.

Uma importante aplicação do agrupamento e reconhecimento de padrões em séries temporais compreende a detecção e o diagnóstico de falhas (FDD, *Fault Detection and Diagnosis*) (MAKI, LOPARO, 1997). A FDD ganhou grande interesse na área industrial em aplicações de monitoramento de condições de máquina. Isso se

deve principalmente às vantagens obtidas através da redução de custos de manutenção e aumento de produtividade (SALAHSHOOR *et al*., 2010). O diagnóstico de falhas em processos (*Process Fault Diagnosis* - PFD) envolve a interpretação do estado normal ou de falha através de medições de variáveis do processo. A detecção prévia das falhas enquanto a planta ainda está operando em uma região controlável pode ajudar a evitar a progressão do evento e, consequentemente, reduzir a possível perda de produção (DASH, VENKATASUBRAMANIAN, 2000). Diferentes técnicas envolvendo agrupamentos têm sido bastante utilizadas em problemas de FDD (LI, WEN, 2014; VENKATASUBRAMANIAN *et al.*, 2003; LEE *et al*., 2004), conjuntamente com técnicas tradicionais (*data driven methods*) (MACGREGOR, CINAR, 2012).

Existem ainda poucos trabalhos sobre o uso de algoritmos genéticos no agrupamento de séries temporais. Dentre estas abordagens encontra-se a de Baragona (2001), a de Liao, *et al*. (2006) e a de Tseng, *et al*. (2009), todos relacionados ao agrupamento de reconhecimento de padrões em séries **univariadas** com diferentes abordagens para a codificação dos cromossomos. Não foram encontrados até o momento trabalhos envolvendo o agrupamento e reconhecimento de padrões em séries temporais multivariadas utilizando algoritmos genéticos. Em todos os trabalhos envolvendo o uso de algoritmos genéticos para agrupamento de séries temporais, independentemente da abordagem adotada para a codificação do cromossomo, as séries consideradas são univariadas. Este trabalho propõe uma abordagem baseada em algoritmo genético e *fuzzy c-means* para agrupamento de séries temporais multivariadas.

O primeiro estudo de caso compreende a detecção de falhas em um processo industrial real. Trata-se do reconhecimento de padrões para detecção de falhas de operação em uma turbina a gás de escala comercial, com capacidade de geração de energia de 27 MW, que representa o principal equipamento da Unidade Termelétrica Rômulo Almeida (UTE-RA, Camaçari-Ba), integrante do parque da Petrobras. Os dados foram obtidos a partir de um sistema de gerenciamento de informações de plantas industriais (*Process Information Management System* – PIMS), disponível na UTE.

O segundo estudo de caso também foi voltado a um problema de FDD e compreendeu o reconhecimento de padrões em eventos normais e com falhas a partir de séries multivariadas extraídas, através de simulação, de uma unidade virtual amplamente utilizada na análise de estratégias de controle e FDD, o *Tennessee Eastman Process* – TEP (DOWNS e VOGEL, 1993; RICKER, 1996).

Na seção seguinte são apresentados o objetivo geral e os específicos deste trabalho. O segundo capítulo traz fundamentos acerca das técnicas aplicadas e destaca trabalhos relacionados ao tema encontrados na literatura. O terceiro capítulo apresenta a metodologia aplicada e os resultados obtidos e os compara com métodos tradicionais de agrupamento. O quarto e último capítulo apresenta as principais conclusões e sugere futuros trabalhos.

1.2 OBJETIVOS

O objetivo geral deste trabalho é demonstrar as potencialidades da aplicação de algoritmo genético no agrupamento de séries temporais multivariadas, tendo os seguintes objetivos específicos:

- Investigar os trabalhos existentes na literatura que tratem do agrupamento de séries temporais por meio de algoritmo genético;
- Obter um banco de dados de séries temporais multivariadas que represente um processo industrial real/virtual;
- Propor uma abordagem de agrupamento de séries temporais multivariadas baseada em algoritmo genético;
- Analisar e implementar alterações no FCM tradicional para ser aplicado a séries temporais multivariadas;
- Comparar o agrupamento multivariado baseado em algoritmo genético com o método clássico FCM.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

2.1 RECONHECIMENTO DE PADRÕES

2.1.1 Visão geral

O reconhecimento de padrões é uma capacidade inerente ao ser humano. A todo momento recebemos dados do mundo à nossa volta através dos nossos sentidos e somos capazes de reconhecer características intrínsecas a estes dados. No geral, fazemos isso rápido e com pouco esforço. Podemos, por exemplo, reconhecer uma pessoa pela sua face, identificar uma voz por telefone e distinguir um perfume específico dentre diversos outros. O reconhecimento de padrões é formalmente definido como o processo pelo qual um padrão recebido é atribuído a uma classe dentre um número predeterminado de classes ou categorias (HAYKIN, 2001). Na maioria dos casos, um padrão representa a descrição numérica ou simbólica de um objeto (ARAÚJO, 2009).

O reconhecimento automático de padrões é utilizado nas mais diversas áreas do conhecimento tais como controle de processos (VENKATASUBRAMANIAN, 2003), biologia (YAO, FREEMAN, 1990), psicologia (OLSHAUSEN *et al*., 1993), medicina (BEZDEK *et al*., 1993), *marketing* (MASUREL *et al*., 2004), visão computacional (AHONEN *et al*., 2006), inteligência artificial (AREL *et al*., 2010) e sensoriamento remoto (MELGANI, BRUZZONE, 2004). Muitos problemas urgentes que surgem na ciência, na indústria e no comércio podem ser tratados como problemas de classificação com um conjunto de dados complexos e, muitas vezes, bastante extensos (MICHIE *et al*., 1994).

Reconhecer padrões, na sua essência, implica na divisão de um conjunto de objetos (ou padrões) em grupos (ou classes) mais ou menos homogêneos com base em

uma medida de similaridade. Objetos semelhantes são alocados em um mesmo grupo (*cluster*) enquanto que objetos que diferem significativamente entre si são alocados em grupos diferentes (PEDRYCZ, 1990). A natureza dos objetos varia de acordo com o tipo de aplicação (sinal de voz, impressão digital, texto, rosto humano, etc...) (OLSZEWSKI, 2001).

De forma geral, um esquema aceito para o processo de reconhecimento de padrões abrange uma sequência de três etapas: aquisição de dados, seleção de atributos e classificação. Na primeira etapa, os dados são coletados do ambiente dentro dos quais os objetos devem ser classificados. Posteriormente, é constituído um espaço de reconhecimento, que envolve a definição de atributos relevantes dos objetos. Finalmente, na terceira etapa, o classificador é construído, ou seja, os objetos são atribuídos a classes baseado em suas características (PEDRYCZ, 1990).

O treinamento do sistema de classificação de objetos pode ser supervisionado ou não supervisionado. No treinamento supervisionado, o padrão de entrada é identificado como membro de um grupo definido previamente. Um objeto é atribuído ao grupo que melhor o representa, baseando-se nos atributos do objeto. Já no treinamento não supervisionado, o padrão é atribuído a um grupo até então desconhecido. Nesse treinamento os grupos são aprendidos com base na similaridade entre os padrões que são iterativamente identificados ao longo da resolução do problema.

As quatro abordagens mais estudadas para o reconhecimento de padrões são: casamento de padrões (*template matching*), classificação estatística, reconhecimento estrutural ou sintático e redes neurais (JAIN *et al*., 2000).

Uma das técnicas mais simples para o reconhecimento de padrões consiste no casamento de padrões. Ele é utilizado para determinar a similaridade entre dois objetos (pontos, curvas ou formas) do mesmo tipo, sendo que um modelo ou um protótipo do padrão a ser reconhecido está disponível. O padrão a ser reconhecido é compatível com o modelo armazenado, considerando aspectos de translação, rotação e escala. A correlação é frequentemente utilizada como medida de similaridade (JAIN *et al*., 2000). Essa técnica possui uma variedade de aplicações, tais como visão computacional, processamento de imagens, biotecnologia, biometria, mineração de dados, entre outras (NOMA, 2010).

Na classificação estatística é determinada a probabilidade de um objeto ser pertencente a uma dada classe (MICHIE *et al*., 1994). Nesta abordagem, cada padrão é

representado como um ponto em um espaço *n*-dimensional. O objetivo dessa representação espacial é estabelecer limites de decisão de forma a separar padrões pertencentes a classes diferentes. O reconhecimento de padrões é realizado em duas etapas: treinamento e teste. O treinamento refere-se a fase de aprendizagem, ou seja, seleção das características apropriadas que descrevam bem os objetos. Nesta fase o classificador é treinado para particionar o espaço amostral. Na fase seguinte (teste) o classificador treinado classifica os objetos não considerados durante a etapa de treinamento, de acordo com as classes de padrões reconhecidas durante a fase de treinamento. A classificação de cada objeto consiste na determinação do grupo ou do padrão com o qual o objeto possui maior similaridade (JAIN *et al*., 2000).

No reconhecimento estrutural de padrões, também conhecido como reconhecimento sintático, grandes conjuntos de dados complexos são descritos por pequenos conjuntos de primitivas de padrão simples e de regras gramaticais. Os padrões são construídos a partir de várias composições de subpadrões, da mesma maneira como as frases são construídas por concatenação de palavras. São definidas várias relações morfológicas entre os subpadrões, que geralmente podem ser expressas em termos de operações lógicas e/ou aritméticas (ALBUS *et al*., 2012). Para que esta abordagem seja vantajosa, os subpadrões - as primitivas do padrão - devem ser mais fáceis de serem reconhecidas do que os próprios padrões (FIRSCHEIN, 1983).

As Redes neurais artificiais (RNAs) são inspiradas nas redes neurais biológicas. As RNAs são extensivamente utilizadas para classificação e agrupamento. Uma rede neural realiza o reconhecimento de padrões passando inicialmente por uma seção de treinamento, durante a qual a rede é apresentada repetidas vezes a um conjunto de padrões de entrada junto com a classe à qual cada padrão específico pertence. Em uma segunda etapa a rede é apresentada a um novo padrão que até então é desconhecido pela rede, mas que pertence a mesma população de padrões utilizada para treiná-la. A rede é então capaz de reconhecer a classe daquele padrão específico devido às informações que a mesma extraiu dos dados de treinamento. Em termos genéricos, as máquinas de reconhecimento de padrões que utilizam redes neurais podem assumir duas formas. Na primeira, a máquina é dividida em duas partes, uma rede não-supervisionada para extração de características e uma rede supervisionada para classificação. Na segunda abordagem, a máquina é projetada como uma única rede de múltiplas camadas. A tarefa de extração das características é realizada pelas unidades computacionais das camadas

da rede utilizando um algoritmo de aprendizagem supervisionado. A escolha entre uma das abordagens vai depender da aplicação de interesse (HAYKIN, 2001). Nesse contexto, a combinação de diferentes métodos de detecção e classificação tem sido uma prática comum no reconhecimento de padrões, já que não existe uma abordagem única considerada ótima para classificação (JAIN *et al*., 2000).

2.1.2 Detecção e diagnóstico de falhas (FDD)

Uma importante aplicação do reconhecimento de padrões consiste na detecção e o diagnóstico de falhas (FDD) (MAKI, LOPARO, 1997). A FDD possui grande importância na área industrial devido às vantagens obtidas a partir da redução de custos de manutenção, melhoria da produtividade e aumento da disponibilidade dos equipamentos (SALAHSHOOR *et al*., 2010).

Para a melhoria da confiabilidade, segurança e eficiência, os métodos avançados de supervisão, detecção e diagnóstico de falhas tornam-se cada vez mais importantes para muitos processos técnicos. Isso é válido especialmente para a segurança de processos e máquinas como aeronaves, trens, automóveis, usinas de energia e plantas químicas (ISERMANN, 2005).

Segundo Isermann e Ballé (1997), uma falha é um desvio não desejado de pelo menos uma propriedade ou parâmetro do sistema da condição aceitável ou padrão. Portanto, a falha é um estado que pode levar a um mal funcionamento ou mesmo parada operacional. Para obter o diagnóstico de falha de processo (PFD), é necessário interpretar o estado atual do processo por meio da análise de suas variáveis. A detecção prévia das falhas, enquanto a planta está trabalhando em uma zona considerada ainda segura, muitas vezes consegue evitar a progressão do evento e, por conseguinte, reduzir prejuízos à produtividade. O PFD constitui o primeiro passo no gerenciamento de situações anormais (*Abnormal Situation Management* - ASM), que visa a detecção, diagnóstico e correção em tempo hábil de condições anormais (DASH, VENKATASUBRAMANIAN, 2000).

Comumente os fabricantes de equipamentos da área industrial fornecem procedimentos de manutenção de forma a inibir ocorrências de falhas, que podem levar a possíveis desligamentos. Os intervalos de tempo entre esses procedimentos recomendados variam de acordo um cronograma de manutenção periódica. Um dos

procedimentos mais caros é a desmontagem, já que apresenta o risco de criar problemas subsequentes como vibração e vazamento. Os sistemas de monitoramento e diagnóstico de falhas são meios eficientes para prevenir manutenções não previstas e onerosas (SALAHSHOOR *et al*., 2010).

De maneira geral, existem quatro procedimentos associados ao monitoramento do processo: detecção de falhas, identificação de falhas, diagnóstico de falhas e recuperação de processos. A detecção de falhas determina se uma falha ocorreu. Se for realizada de maneira precoce pode fornecer um aviso importante sobre problemas emergentes, de forma que ações apropriadas podem ser tomadas para evitar sérios transtornos do processo. A identificação de falhas aponta as variáveis de observação mais relevantes para o diagnóstico da falha, de modo que o efeito da falha pode ser eliminado de forma mais eficiente. O diagnóstico de falha determina qual a falha ocorrida e o tipo, localização, magnitude e tempo da falha. A recuperação do processo, também chamada de intervenção, remove o efeito da falha e é o procedimento necessário para fechar o *loop* de monitoramento do processo (Figura 1).

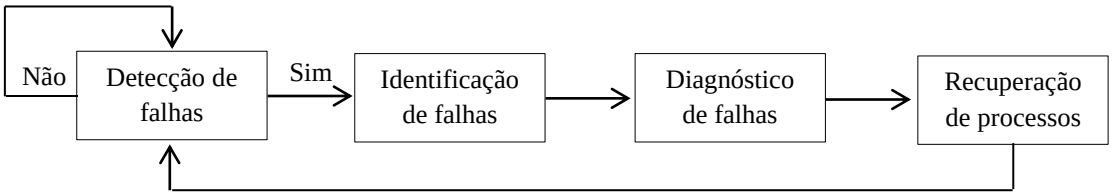


Figura 1 – Um esquema do *loop* de monitoramento de processo (CHIANG, *et al*., 2000).

A área de detecção e diagnóstico de falhas é um importante aspecto da engenharia de processo. É importante não só do ponto de vista da segurança, mas também do impacto econômico envolvido. A detecção rápida e o diagnóstico de falhas são essenciais para a operação confiável, segura e eficiente da planta e para manter a especificação dos produtos. Em um ambiente industrial, podem ocorrer falhas no processo e nos equipamentos de forma independente ou simultânea. Em sistemas industriais complexos, geralmente é muito difícil medir diretamente estados de processo que são bons indicadores de falhas, para isso são necessárias medidas mais elaboradas e automáticas. Muitas vezes, observando dados de múltiplas variáveis ao mesmo tempo, operadores qualificados são necessários para tomar decisões difíceis com base em suas experiências e conhecimento empírico. Sistemas inteligentes de suporte a operadores

em tempo real são vistos como uma alternativa para apoio à tomada de decisão (MAKI, LOPARO, 1997).

Em geral, os métodos de classificação ou reconhecimento de padrões para FDD compreendem os que são baseados em um modelo do processo e os que extraem conhecimento através de dados históricos do processo (DASH, VENKATASUBRAMANIAN, 2000).

2.2 ANÁLISE DE AGRUPAMENTOS

2.2.1 Visão geral

O avanço da tecnologia da informação viabilizou o desenvolvimento de métodos associados à extração de conhecimento presente em dados. A partir da mineração de dados é possível analisar grandes conjuntos de dados para buscar conhecimento útil (KANTARDZIC, 2011, ELMASRI; NAVATHE, 2005). Uma das abordagens de mineração de dados consiste em classificá-los ou agrupá-los em categorias conforme o grau de similaridade.

A classificação, uma das atividades mais primitivas dos seres humanos, desempenha um papel importante e indispensável na história do desenvolvimento humano. Para aprender sobre um novo objeto ou entender um fenômeno novo, as pessoas sempre tentam buscar características (atributos) que podem descrevê-lo e compará-lo com outros objetos ou fenômenos já conhecidos. Ao fazermos isso, estamos seguindo alguns padrões ou regras com base na semelhança ou diferença (XU, WUNSCH, 2005).

Na análise de agrupamentos, os objetos são divididos em grupos mais ou menos homogêneos, com base em uma medida de similaridade (XU, WUNSCH, 2005). O objetivo do processo é criar grupos de tal forma que elementos pertencentes ao mesmo grupo sejam muito similares, enquanto que elementos de grupos diferentes sejam bastante dissimilares. Os grupos reconhecidos representam comportamentos distintos (OLIVEIRA *et al*., 2007).

Como ilustração, a Figura 2 apresenta uma série de objetos representados por duas características. Trata-se de um conjunto de 20 cidades, onde cada uma é caracterizada por suas coordenadas geográficas (latitude e longitude). Uma técnica de

agrupamento deve ser capaz de reconhecer dois grupos e encontrar dois centros no conjunto de cidades para que duas novas antenas sejam instaladas por uma empresa de telefonia celular (classificação representada pelas cores vermelho e azul com seus respectivos centros). Os centros (padrões) obtidos representam a localização ótima para a instalação de apenas duas antenas de modo a atender ao conjunto dos 20 municípios (OLIVEIRA *et al*., 2007).

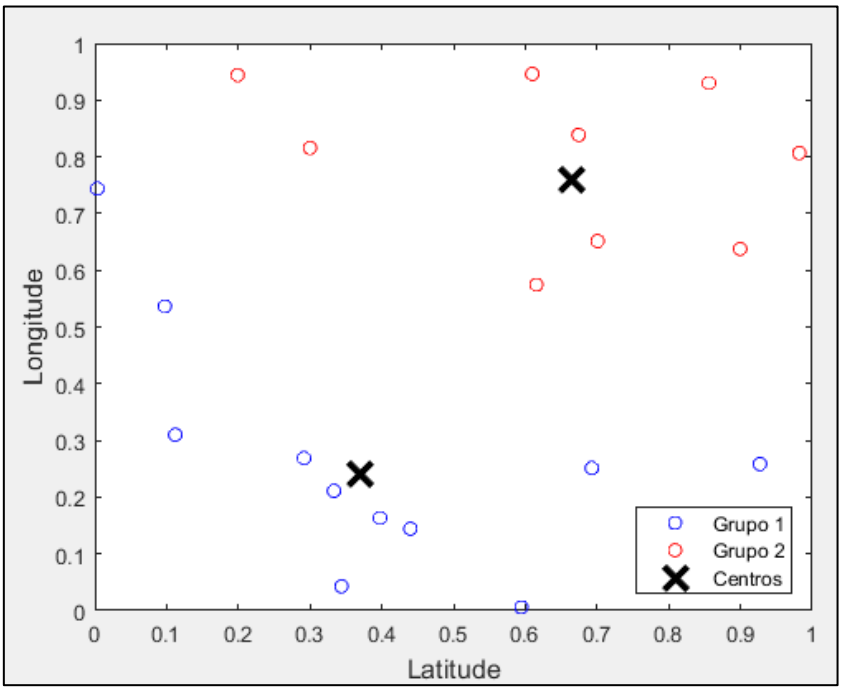


Figura 2 – Agrupamento 20 cidades em dois grupos.

No contexto de reconhecimento de padrões, a análise de agrupamentos está dentro das técnicas de reconhecimento de padrões estatístico. Como um ramo da estatística, a análise de agrupamentos foi amplamente estudada, com o foco principal na análise de grupo baseada em distância (HAN *et al*., 2011). Na análise de agrupamentos não se conhece previamente a informação do rótulo da classe associada a cada objeto. Portanto, neste caso não há informação que permita a distinção prévia dos objetos a serem agrupados. Por esse motivo, o agrupamento é uma forma de aprendizagem por observação e não por meio de exemplos. Além disso, o número de categorias, classes ou grupos pode não ser previamente conhecido.

Em problemas mais simples, a própria percepção humana é capaz de realizar o agrupamento de maneira satisfatória, independente da utilização de técnicas de análise de agrupamentos. Em problemas complexos, é imprescindível a aplicação de técnicas

ou algoritmos específicos para o reconhecimento de grupos e padrões. (HAN *et al*., 2011).

2.2.2 Principais etapas

O processo de agrupamento de padrões envolve pelo menos três etapas principais: (1) seleção/extração de características, (2) definição de uma métrica de similaridade (3) agrupamento propriamente dito (JAIN *et al*., 1999).

A seleção de características é o processo de identificação do subconjunto mais relevante das características ou atributos originais a serem utilizadas e a extração de características compreende a criação de novas características a partir de transformações ou combinações do conjunto de características original (XU, WUNSCH, 2005). A existência de recursos irrelevantes no conjunto de dados pode reduzir a qualidade da aprendizagem e consumir mais tempo e memória computacional. Uma boa seleção de características pode diminuir significativamente a carga de trabalho e simplificar o processo de aprendizado (AGGARWAL, REDDY, 2013).

O grau de dissimilaridade entre os objetos geralmente é medido por uma função de distância. Uma medida de distância deve ser simétrica e o seu valor mínimo (geralmente zero) implica em total identidade entre os objetos (ROKACH, 2009). O agrupamento baseado em otimização combina a escolha de uma medida de similaridade e a definição de uma função objetivo. Os algoritmos de agrupamento estão explicitamente ou implicitamente conectados à definição da métrica de similaridade.

A etapa de agrupamento pode ser realizada de diversas maneiras, mas geralmente é combinada com a escolha de uma medida de similaridade correspondente e com a construção de uma função objetivo. Uma vez escolhida uma medida de similaridade, a construção de uma função objetivo de agrupamento torna a partição de grupos um problema de otimização, que está bem definido matematicamente e possui ótimas soluções na literatura (XU, WUNSCH, 2005).

2.2.3 Características

Um bom algoritmo de agrupamento deve possuir bons resultados com amostras de treinamento e com amostras de teste. As amostras denominadas de treinamento são

aquelas utilizadas para o treinamento do algoritmo, independente da regra de classificação ou decisão usada. O desempenho do classificador depende da quantidade e dos valores específicos das amostras de treinamento disponíveis. Em paralelo, o objetivo de projetar um sistema de reconhecimento de padrões é classificar futuras amostras de teste que geralmente são distintas das amostras de treinamento. Classificar bem padrões de teste que não foram utilizados durante a etapa de treinamento implica em validar a capacidade de generalização do classificador. Destarte, otimizar o desempenho de um classificador no conjunto de treinamento nem sempre resulta em um ótimo desempenho em um conjunto de teste. Uma fraca capacidade de generalização de um classificador pode ser atribuída a um dos seguintes fatores: o número de características é muito grande em relação ao número de amostras de treinamento, o número de parâmetros desconhecidos associado ao classificador é grande ou o classificador aprendeu de forma excessiva o conjunto de treinamento (*overtrained*) (JAIN *et al*., 2000).

Um algoritmo de agrupamento sempre pode gerar uma divisão em um conjunto de dados, independente da existência ou não da estrutura. Além disso, diferentes abordagens geralmente levam a diferentes grupos. Para o mesmo algoritmo, a identificação dos parâmetros ou a ordem de apresentação dos padrões de entrada podem afetar os resultados finais. Muitas vezes pode ser necessário que especialistas na área interpretem a partição de dados, de forma a garantir a confiabilidade do conhecimento extraído (XU, WUNSCH, 2005).

Os métodos de agrupamento são categorizados em métodos de particionamento (não hierárquicos), hierárquicos e baseados em densidade (AGGARWAL, REDDY, 2013). Dado um conjunto de n objetos não rotulados, um método de particionamento (não hierárquico) determina k partições de dados e cada partição representa um grupo que deve conter pelo menos um objeto ($k \leq n$). A partição é rígida se cada objeto pertence completamente e exclusivamente a um dado grupo. A partição é *fuzzy* se for permitido a um objeto pertencer a mais de um grupo com diferentes graus de intensidade. Dois métodos bastante populares de particionamento rígido são o *k-means*, onde o padrão ou centro de cada grupo é representado pelo valor médio de seus objetos e o *k-medoids* onde o centro de cada grupo é representado por seu objeto mais centralizado (LIAO, 2005).

Um método de agrupamento hierárquico faz uma hierarquia de grupos usando algoritmos aglomerativos ou divisivos. Os métodos aglomerativos iniciam colocando cada objeto em seu próprio grupo e, em seguida, mesclam os grupos similares. Assim, os grupos são formados reunindo-se os objetos em grupos cada vez maiores. É possível verificar as fusões sucessivas dos indivíduos até que todos os objetos estejam em um único grupo ou até que certas condições de término sejam satisfeitas como o número desejado de grupos. Os métodos divisivos fazem o oposto. Todos os objetos partem de um grande grupo e estes são subdivididos em dois subgrupos de tal forma que os objetos dos mesmos subgrupos tenham o máximo de semelhança e elementos de subgrupos distintos tenham grande dissimilaridade. Esses subgrupos são posteriormente subdivididos em outros subgrupos (LIAO, 2005).

O agrupamento baseado em densidade tem como objetivo determinar grupos de alta densidade de objetos separados por regiões de baixa densidade. Este agrupamento tenta explorar o espaço de dados em níveis elevados de granularidade baseado em densidade de registros por região, o que permite a formação de grupos com formatos arbitrários. Assim, os grupos formados crescem de acordo com a densidade de dados em um "potencial" grupo. O principal desafio dos métodos baseados em densidade é que eles são naturalmente definidos em pontos de dados em um espaço contínuo, o que restringe a sua aplicação de um espaço amostral discreto. Assim, muitos tipos de dados arbitrários, como dados de séries temporais, não são tão fáceis de usar com métodos baseados em densidade sem transformações especializadas (AGGARWAL, REDDY, 2013).

2.2.4 O agrupamento Fuzzy

O modelo de agrupamento *fuzzy* foi introduzido na literatura por Dunn em 1973 e ampliado por Bezdek em 1981. A teoria de agrupamento *fuzzy* representa uma poderosa ferramenta para resolução de problemas de agrupamento, diferenciando-se da abordagem rígida na qual cada objeto pertence única e exclusivamente a um determinado grupo (DELGADO *et al*., 1997). A teoria inicial de conjuntos *fuzzy* foi proposta por Zadeh em 1965. Essa teoria trouxe a ideia de incerteza na associação de um objeto a um dado conjunto, que foi descrita por uma função de pertinência. A partição de objetos em diferentes grupos possui uma incerteza intrínseca na medida em

que um objeto pode pertencer simultaneamente a mais de um grupo com diferentes graus de intensidade e os grupos possuem sobreposição entre si, não havendo uma fronteira claramente definida entre eles. A aplicação da teoria de Zadeh na análise de agrupamentos foi proposta inicialmente no trabalho de Bellman, Kalaba e do próprio Zadeh em 1966. A partir de então, o agrupamento *fuzzy* foi amplamente estudado e aplicado em uma grande variedade de áreas do conhecimento (YANG, 1993). Classificação de padrões, segmentação de imagens, categorização de documentos, identificação de sistemas dinâmicos, por exemplo, utilizam agrupamento *fuzzy*, (HRUSCHKA et al., 2009).

Para ilustrar a essência do agrupamento *fuzzy*, considere um objeto que seja equidistante de cinco grupos. Um algoritmo rígido (abordagem tradicional) teria que escolher um dos grupos para a inevitável atribuição de pertinência. Mas qual deles escolher? Afinal, em relação ao objeto os grupos são indistintos pelo critério da distância. Em outras palavras, os dados muitas vezes compreendem categorias que se sobrepõem umas às outras em algum grau. Nesse contexto, o agrupamento *fuzzy* trabalha com a ideia de que se os pontos são equidistantes de cinco grupos, eles "pertencem" igualmente a cada um deles, ou seja, a pertinência do ponto em relação a cada um dos grupos é de 0,2 (OLIVEIRA *et al*., 2007).

Em 1981, Bezdek propôs o algoritmo *fuzzy c-means* (FCM) que consiste no primeiro método eficiente de partição *fuzzy* (DE OLIVEIRA, PEDRYCZ, 2007). O FCM é capaz de resolver problemas de agrupamento nos quais a estrutura ou definição dos grupos não se apresenta de forma tão evidente e há sobreposição entre as fronteiras dos grupos (NAYAK *et al*., 2015; MELIN, CASTILLO, 2014; BEZDEK *et al*., 2006).

A Figura 3 ilustra uma partição rígida (retângulos R1 e R2) e *fuzzy* (elipses F1 e F2) envolvendo a mesma amostra de dados.

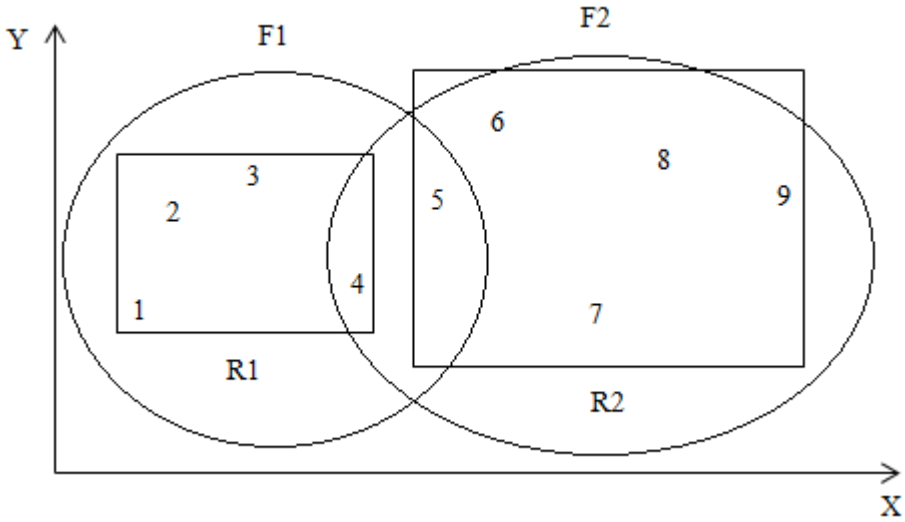


Figura 3 – Agrupamentos rígido e *fuzzy* (Adaptado de JAIN *et al*., 1999).

2.2.4.1 Fuzzy c-means

O *fuzzy c-means* é baseado no conceito de que cada objeto não deve pertencer exclusivamente a um determinado grupo (BEZDEK *et al*., 1984). Nesse algoritmo cada objeto pertence a diversos grupos com graus de adesão específicos (entre 0, pertinência nula e 1, pertinência total). Os graus de pertinência são dispostos em uma matriz de pertinência (HRUSCHKA *et al*., 2009).

O algoritmo *fuzzy c-means* é a extensão (para o domínio difuso) do algoritmo *k-means* (MACQUEEN, 1967; JAIN, DUBES, 1988; JAIN, 2010). O FCM clássico é formulado através de um problema de otimização que consiste na minimização do somatório das distâncias de cada objeto aos centros dos grupos, ponderado pela respectiva pertinência do objeto a cada grupo (LIAO, 2005):

$$\min \underbrace{J}_{U,V} = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^{m} \cdot \|x_k - v_i\|^2 \qquad (1)$$

onde N é o número de objetos, C é o número de grupos, $X = \{x_1, x_2, \ldots, x_N\}$ é a amostra de N objetos (de treinamento), $V = \{v_1, v_2, \ldots, v_C\}$ é o vetor com os centros (padrões) dos grupos, u_{ik} é o grau de pertinência do objeto x_k ao grupo i e m é o parâmetro fuzzificador que controla o grau de sobreposição entre os grupos obtidos ($m > 1$ e em geral $m = 2$).

O problema de otimização envolve ainda as seguintes restrições:

$$0 \leq u_{ik} \leq 1 \tag{2}$$

$$\sum_{i=1}^{C} u_{ik} = 1, \ \ \forall k. \tag{3}$$

$$\sum_{k=1}^{N} u_{ik} > 0, \ \ \forall i. \tag{4}$$

O parâmetro *m* é importante no que se refere ao grau permitido de "mistura" dos grupos, também chamado de grau de fuzzificação do agrupamento. Se m for muito próximo de 1, as partições obtidas ficarão próximas da abordagem de partição rígida (graus de pertinência tenderão a 1 ou zero).

O operador de norma $\|.\|$ representa a métrica de similaridade adotada. Em geral, quando cada objeto é representado por um ponto no $\Re^n$ (caso de séries temporais univariadas), a distância Euclidiana é a métrica mais adotada (HRUSCHKA *et al*., 2009).

A primeira restrição de desigualdade, equação (2), estabelece que o grau de pertinência está restrito ao intervalo [0,1]. A segunda, equação (3), determina que a soma das pertinências de um dado objeto a todos os grupos deve ser igual a 1 (abordagem probabilística de agrupamento *fuzzy*) e a terceira restrição, equação (4), simplesmente impede a formação de grupos vazios.

O problema de minimização apresentado (eq.s 1-4) possui a matriz de pertinência $(u_{ik}, i = 1, \dots, C \ e \ k = 1, \dots N)$ e os centros de cada grupo $(v_i, i = 1, \dots, C)$ como variáveis de decisão. A função objetivo estabelece que objetos mais distantes dos centros tenham graus de pertinência necessariamente menores, o que estabelece uma relação de interdependência entre os centros de cada grupo e os níveis de pertinência dos objetos aos grupos.

A aplicação das condições de primeira ordem ao problema de otimização definido pelas eq.s (1) a (4) produz como resultado as seguintes equações que inter-relacionam as variáveis de decisão do problema (graus de pertinência e centros/padrões dos grupos):

$$u_{ik} = \frac{\left(\frac{1}{\left\| x_k - v_i \right\|^2} \right)^{\frac{1}{m-1}}}{\sum_{j=1}^{c} \left(\frac{1}{\left\| x_k - v_j \right\|^2} \right)^{\frac{1}{m-1}}} \quad (5)$$

$$v_k = \frac{\sum_{k=1}^{n} u_{ik}^{\ m} \cdot x_k}{\sum_{k=1}^{n} u_{ik}^{\ m}} \quad (6)$$

A equação (5) caracteriza o problema de partição *fuzzy* probabilística. O grau de pertinência de um objeto a um dado grupo não depende apenas da distância do objeto a este grupo, mas também das distâncias deste mesmo objeto aos demais grupos.

O algoritmo FCM clássico compreende a aplicação das eq.s (5) e (6) de forma sucessiva (otimização alternada). Este processo é realizado de modo iterativo tendo-se a atualização da matriz de pertinência, equação (5), e dos centros de grupo, equação (6), a cada iteração, até que um critério de parada/convergência seja satisfeito (OLIVEIRA *et al.*, 2007). Neste caso, é necessário uma estimativa inicial para os centros ou para as pertinências dos objetos.

O *fuzzy c-means* é conhecido como um método de classificação estável e robusto. Devido à sua simplicidade e baixas demandas computacionais, o FCM é amplamente utilizado como base para a proposição de novas abordagens de agrupamento (DE OLIVEIRA, PEDRYCZ, 2007).

2.3 SÉRIES TEMPORAIS

As séries temporais representam um objeto essencialmente dinâmico que pode ser agrupado (LIAO, 2005). Uma série temporal é um conjunto de observações feitas sequencialmente ao longo do tempo (ELMASRI; NAVATHE, 2005). É bastante comum que sejam representados por séries temporais dados de sensores, economia, *marketing* ou qualquer outro tipo de aplicação de rastreamento ou previsão temporal (AGGARWAL, REDDY, 2013). Dados históricos estão em geral associados a um

comportamento dinâmico e ao conceito de causalidade que compreende a dependência do estado atual de um sistema em relação aos estados passados (SCHKODA, 2012).

O aspecto principal das séries temporais é que os valores instantâneos dos dados não são independentes uns dos outros. As séries contêm um atributo contextual (tempo) e um atributo comportamental (valor dos dados) (AGGARWAL, REDDY, 2013). A Figura 4 ilustra uma série temporal, que representa o percentual de pessoas de 5 anos ou mais de idade alfabetizadas no Brasil no período de 2001 a 2011.

Figura 4 – Pessoas alfabetizadas de 5 anos ou mais de idade no Brasil.
Fonte: IBGE, Pesquisa Nacional por Amostra de Domicílio 2001-2011.

O caso mais simples compreende a série temporal univariada quando cada objeto contém uma única série temporal referente a uma determinada variável de processo. Séries temporais multivariadas compreendem conjuntos de séries associadas a diferentes variáveis de processo (LIAO, 2005). Segundo Aggarwal e Reddy, 2013, processos multivariados surgem quando vários processos de séries temporais relacionadas são observados simultaneamente ao longo do tempo e não apenas uma única série. No estudo de séries multivariadas, é necessária uma abordagem que descreva não apenas as propriedades de cada série, mas também o comportamento integrado entre elas.

Genericamente, dada a seguinte matriz tridimensional:

$$\boldsymbol{X} \equiv \{x_{kjt} : k = 1, \dots, N;\ j = 1, \dots, J; t = 1, \dots, T\}, \quad (7)$$

um objeto *k* pode ser representado pela seguinte matriz:

$$\boldsymbol{X}_k = \begin{bmatrix} x_{k11} & x_{k21} & \cdots & x_{kJ1} \\ x_{k12} & x_{k22} & \cdots & x_{kJ2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1T} & x_{k2T} & \cdots & x_{kJT} \end{bmatrix} \quad (8)$$

onde k representa o objeto, j a variável e t o instante de tempo. Desta forma, x_{kjt} representa o valor da variável j no objeto k (X_k), observado (ou medido) no instante de tempo t. Quando $J = 1$, cada objeto é uma série temporal univariada, quando $J \geq 2$ tem-se o caso multivariado. Cada coluna, portanto, contém a série temporal relacionada à variável j no objeto *k*. A matriz tridimensional $\boldsymbol{X}$ pode ilustrada pelo sólido da Figura 5. Os eixos horizontal, vertical e de profundidade são representados pelos índices k, j e t, repectivamente.

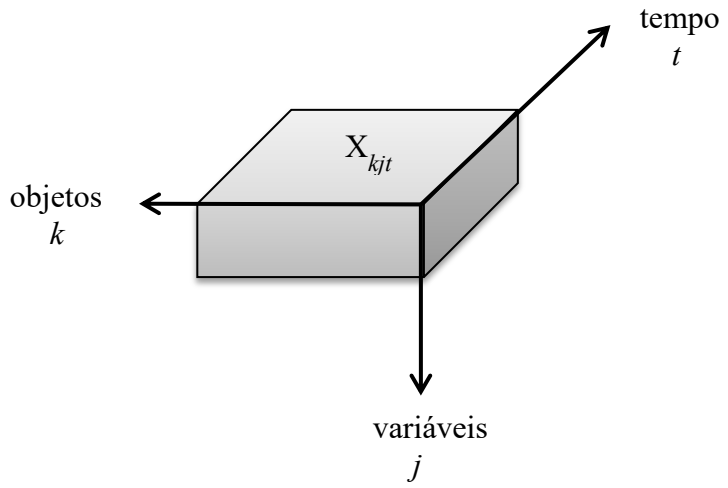


Figura 5 – Representação de séries temporais multivariadas.
Fonte: Adaptado de D'URSO e MAHARAJ, 2012.

2.3.1 Agrupamento e Reconhecimento de Padrões em Séries Temporais

Cada série temporal pode ser analisada como um único objeto no contexto de um problema de agrupamento (WANG, *et al.*, 2006). O agrupamento de séries temporais pode trabalhar diretamente com dados brutos, indiretamente com características extraídas ou com os modelos identificados a partir de cada série (RANI, SIKKA, 2012).

Considera-se duas abordagens no agrupamento de séries temporais (Keogh e Lin, 2005). A primeira é o agrupamento inteiro ou total (*whole clustering*) que consiste na aplicação de um método tradicional de agrupamento (hierárquico ou não) considerando as séries como objetos. A segunda é o agrupamento instantâneo ou subsequencial (*subsequence clustering*) que consiste em utilizar como amostra uma única série temporal e extrair desta séries menores (subsequências) através de uma janela deslizante que contém um padrão pré-definido (problema de segmentação).

Para a escolha do algoritmo de agrupamento, alguns fatores devem ser observados tais como o período de amostragem, natureza das séries (univariadas ou multivariadas) e uniformidade (ou não) entre as janelas de tempo dos objetos. Algoritmos clássicos de partição não hierárquicos, tais como o FCM, podem ser utilizados para o agrupamento de séries temporais considerando neste caso cada série como um objeto cujas características serão necessárias para quantificar a similaridade em relação a candidatos a centros dos respectivos grupos.

A distância Euclidiana é a métrica de similaridade mais utilizada para as séries univariadas (XUN, ZHISHU, 2010; WANG, *et al*., 2013). Apesar de ser aplicável também para séries multivariadas, neste caso a distância Euclidiana não permite uma análise do comportamento global ou integrado das séries que constituem o objeto (BANKÓ, ABONYI, 2012). Portanto, são imprescindíveis ferramentas estatísticas que apresentem uma visão mais global do fenômeno, não se limitando a observar as variáveis (cada série) de modo isolado.

A análise de componentes principais – PCA (do inglês *Principal Component Analysis*) oferece uma importante alternativa para a quantificação de similaridade entre séries multivariadas (MAGNUSSON, 2003). Esta técnica possibilita investigações por meio da extração de características associadas ao comportamento conjugado das séries que constituem o objeto. Nesse contexto, vários trabalhos relatam aplicações bem sucedidas envolvendo métrica baseada em PCA (ROSÉN, YUAN, 2001; HUANG *et al*., 2000; MANSFIELD *et al*., 1997; LIN *et al*., 1997).

No PCA, um conjunto original de variáveis correlacionadas é transformado em um conjunto de variáveis descorrelacionadas: os componentes principais (PC). Essa transformação busca eliminar algumas variáveis originais que possuam poucas informações, preservando, ao máximo, a variabilidade do conjunto. Se existirem *p* variáveis originais, então existirão no máximo *p* componentes principais. Entretanto, a

redução do número de variáveis não se faz por uma simples seleção de algumas variáveis, mas pela construção de novas variáveis sintéticas, obtidas pela combinação linear das variáveis inicias. Nessa transformação, ordena-se os componentes principais de modo que o primeiro deles esteja relacionado à maior variabilidade dos dados contidos na matriz (ALAEI *et al*., 2013).

Considere a seguinte matriz de dados:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \cdots & x_{kj} \end{bmatrix} \quad (9)$$

onde j é o número de variáveis medidas e k é o número de pontos ou medições de cada variável. As variáveis X_1, X_2, X_3, ..., X_j representam as colunas na matriz *X*. O PCA gera novas variáveis, Z_1, Z_2, Z_3, ..., Z_j, , não correlacionados e capazes de preservar a variabilidade original dos dados (matriz *X*) (VARELLA, 2008). Para isso, esta técnica baseia-se na matriz de covariância dos dados, de onde são extraídos os autovalores e os autovetores (JOLLIFFE, 2002). A covariância está associada à variabilidade entre dois vetores (eq. 10):

$$Cov(X_1, X_2) = \frac{\sum_{t=1}^{k}[(X_1(t) - \overline{X}_1) \cdot (X_2(t) - \overline{X}_2)]}{k} \quad (10)$$

onde X_1 e X_2 são as duas variáveis analisadas, $X_1(t)$ e $X_2(t)$ são as respectivas medições a cada instante de tempo e $\overline{X}_1$ e $\overline{X}_2$ são as médias.

A matriz de covariância amostral (C) é simétrica e composta por todas as covariâncias entre as variáveis (colunas) que compõem a amostra original (colunas da matriz X).

$$C = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) & \cdots & Cov(X_1,X_j) \\ Cov(X_2,X_1) & Var(X_2) & Cov(X_2,X_3) & \cdots & Cov(X_2,X_j) \\ Cov(X_3,X_1) & Cov(X_3,X_2) & Var(X_3) & \cdots & Cov(X_3,X_j) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_l,X_1) & Cov(X_l,X_2) & Cov(X_l,X_3) & \cdots & Var(X_j) \end{bmatrix} \quad (11)$$

O PCA consiste em obter os autovalores e autovetores da matriz de covariância. As variâncias dos componentes principais estão diretamente relacionadas aos autovalores da matriz C. Assumindo-se que os autovalores estão ordenados como $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_j \geq 0$, então λ_i corresponde a variância do i-ésimo componente principal Z_i.

Uma propriedade importante dos autovalores é que a soma deles é igual à soma dos elementos da diagonal da matriz C, eq. 12.

$$\lambda_1 + \lambda_2 + \cdots + \lambda_j = Var(X_1) + Var(X_2) + \cdots + Var(X_j) \quad (12)$$

isso significa que a soma das variâncias dos componentes principais é igual a soma das variâncias das variáveis originais. Portanto, os componentes principais abrangem a variabilidade dos dados originais (MANLY, 2008).

O PCA fornece uma eficiente métrica de similaridade entre séries temporais multivariadas denominada SPCA (PCA *Similarity Factor*). Havendo duas STM (dois objetos) contendo o mesmo número de variáveis (mas não necessariamente o mesmo número de medições), o SPCA fornece um único índice para quantificar a similaridade entre estes dois objetos através da comparação entre as direções de seus respectivos componentes principais (LI, WEN, 2014; DENG, TIAN, 2013). Um valor de SPCA próximo de 1 indica que os objetos são bastante similares e um valor próximo de zero indica uma elevada dissimilaridade.

$$SPCA = \frac{1}{k} \cdot \sum_{i=1}^{k} \sum_{j=1}^{k} cos^2 \theta_{ij} \quad (13)$$

onde k é o número de componentes principais (PCs) selecionados em ambos os objetos (X_1e X_2) e θ_{ij} é o ângulo entre o *i*-ésimo PC de X_1 e o *j*-ésimo PC de X_2. A divisão por k visa apenas normalizar o valor SPCA. O número k de PCs é determinado a partir da quantidade necessária para descrever 95% da variabilidade total em cada conjunto de

dados originais. O índice SPCA mede a semelhança entre duas matrizes (dois objetos) ao computar o cosseno dos ângulos de todas as combinações entre componentes principais das duas matrizes (SINGHAL, SEBORG, 2005).

No SPCA original (eq. 13) todos os PCs possuem o mesmo peso, o que não torna possível a captura do grau de similaridade entre os conjuntos de dados quando apenas um ou dois PCs explicam a variância. Nesse contexto, um fator de similaridade modificado, S_{PCA}^{λ}, é definido de forma a atribuir peso aos componentes principais de acordo com o autovalor a este associado (SINGHAL e SEBORG, 2005; DENG e TIAN, 2007; FONTES e BUDMAN, 2018).

$$S_{PCA}^{\lambda} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\lambda_i^{(1)}\,\lambda_j^{(2)}\right)cos^2\theta_{ij}}{\sum_{i=1}^{k}\lambda_i^{(1)}\,\lambda_i^{(2)}} \tag{14}$$

onde $\lambda_i^{(1)}$ e $\lambda_i^{(2)}$ são os *i*-ésimos autovalores do primeiro e segundo objetos, respectivamente.

2.4 ALGORITMO GENÉTICO

Tentando superar algumas limitações associadas aos algoritmos clássicos de otimização, versões modificadas dos algoritmos clássicos de agrupamento (tais como o próprio FCM) têm sido propostas baseadas no uso de algoritmos genéticos (GAs).

Os algoritmos genéticos fazem parte de uma família de modelos computacionais inspirados pela evolução natural. Esses algoritmos codificam uma potencial solução para um problema específico em uma estrutura de dados denominada de cromossomo. Em um uso mais amplo do termo, um algoritmo genético é qualquer modelo baseado em população que utiliza seleção e operadores de recombinação para gerar novas possíveis soluções em um espaço de busca. Embora a gama de problemas aos quais estes algoritmos foram aplicados seja bastante ampla, uma aplicação comum dos GAs é como otimizadores de funções (WHITLEY, 1994).

Os algoritmos genéticos procuram o espaço de solução de uma função através do uso da evolução simulada no qual o indivíduo mais capacitado deve sobreviver. Esses algoritmos foram desenvolvidos para resolver problemas lineares e não lineares, executando uma pesquisa probabilística em um espaço grande de busca (HOUCK *et al.*,

1995). O GA tem se mostrado como um algoritmo eficaz de tentativa de otimização global (MENG *et al*., 2002). Uma simplificação dos algoritmos genéticos pode ser vista na Figura 6.

(1) Fornecer uma população P_0 de N indivíduos e seus respectivos valores de função
(2) $i \leftarrow 1$
(3) $P_i' \leftarrow$ função_de_selecao$(P_i - 1)$
(4) $P_i \leftarrow função\_de\_reprodução(P_i')$
(5) $Avalie(P_i)$
(6) $i \leftarrow \mathrm{i} + 1$
(7) Repita a partir do passo 3 até o término
(8) Imprimir a melhor solução encontrada

Figura 6 – Um algoritmo genético simples (HOUCK *et al*., 1995).

O uso de um algoritmo genético requer a definição de seis aspectos fundamentais: representação cromossômica, criação da população inicial, função de avaliação, função de seleção, operadores genéticos que compõem a função de reprodução e critérios de término (HOUCK *et al*., 1995). Nas próximas sessões essas questões serão descritas.

2.4.1 Representação dos Cromossomos

Para qualquer algoritmo genético, é necessária uma representação cromossômica para descrever cada indivíduo na população. A forma de representação determina como o problema será estruturado. Um cromossomo (ou indivíduo) é formado por uma coleção de genes que representam parâmetros ou variáveis de decisão. Estes genes pertencem a um determinado alfabeto, que pode consistir em dígitos binários (0 e 1), números inteiros, números reais, letras, entre outros. Inicialmente, o alfabeto estava limitado a dígitos binários. Demonstrou-se que as representações mais próximas da realidade são mais eficientes e produzem melhores soluções (MICHALEWICZ, 1994). Uma representação útil de um indivíduo para otimização de função envolve genes com número reais dentro de limites superiores e inferiores. Se comparados GAs de valor real e binário, o de valor real é mais eficiente em termos de tempo de CPU. Além disso, uma representação de valor real está mais próxima da representação do problema, o que

oferece maior precisão com resultados mais consistentes (HOUCK *et al*., 1995). Diante do exposto, a codificação real para os cromossomos foi a escolhida para este trabalho.

2.4.2 População Inicial, Término e Função de Avaliação

A população de soluções é formada por soluções candidatas para o problema, que estão codificadas como cromossomos. O que diferencia uma solução de outra é a aptidão, ou seja, a qualidade ou capacidade da solução para resolver o problema em questão. Um algoritmo genético inicia criando uma população inicial de cromossomos (passo 1 da Figura 6) e realiza a evolução em direção a uma solução ótima por meio de gerações. Cada iteração completa do algoritmo constitui uma geração. Assim, a população inicial é considerada a primeira geração. O mais comum é a população inicial ser escolhida de maneira aleatória (HOUCK *et al*., 1995).

As possíveis soluções se movem de geração em geração até que um critério de término seja atendido, como, por exemplo, a soma dos desvios entre os indivíduos se torna desprezível e/ou um número máximo de gerações consecutivas sem alteração da solução é alcançado (CHIOU, LAN, 2001).

As funções de avaliação, também conhecidas como funções *fitness*, podem ser usadas de várias formas em um algoritmo genético. Entretanto, o requisito mínimo para esta função é o mapeamento da população em um conjunto parcialmente ordenado por meio de um critério de qualidade (valor de aptidão) (HOUCK *et al*., 1995). Ao ordenar os indivíduos ou soluções, a função de avaliação faz o papel do ambiente na seleção natural. Essa função avalia cada cromossomo segundo uma métrica de avaliação da qualidade de forma que se atribui uma nota de desempenho a cada um deles. O valor de aptidão é utilizado para orientar a seleção dos indivíduos mais aptos que terão maior probabilidade de propagar sua informação genética, ou seja, as melhores soluções serão utilizadas para gerar novas soluções ainda melhores (BANDYOPADHYAY, MAULIK, 2002).

2.4.3 Operadores Genéticos

Os operadores genéticos fornecem o mecanismo de pesquisa básico do GA. Operadores genéticos geram e alteram a composição da prole preservando, porém,

algumas características essenciais. Existem três destes operadores: seleção, *crossover* e mutação. Especificadamente os operadores de reprodução de *crossover* e mutação são usados para criar novas soluções baseadas em soluções já existentes na população (HOUCK *et al*., 1995).

2.4.3.1 Seleção

Um operador de extrema importância para o algoritmo genético, o operador de seleção faz a seleção probabilística com base nos valores de aptidão, de modo que os indivíduos mais aptos tenham maior chance de serem selecionados. Os cromossomos selecionados são os melhores para produzir a descendência. Dessa forma, esse operador age como a seleção natural, que se encarrega de manter os melhores indivíduos na população (BARAGONA, 2001).

A cada geração, após a adaptação de cada cromossomo na população ser medida pela função de avaliação, alguns indivíduos são selecionados para a próxima geração. A operação de seleção escolhe os *p* melhores indivíduos dentre a população original de tamanho *p*, de tal maneira que a próxima geração sempre possui o mesmo tamanho da anterior. Portanto, um mesmo indivíduo pode ser selecionado mais de uma vez (indivíduos com maior aptidão), assim como todos os indivíduos da população também têm a chance de serem selecionados para serem reproduzidos na próxima geração (HOUCK *et al*., 1995).

Existem vários esquemas para o processo de seleção: seleção roleta (*roulette wheel selection*) e suas extensões, técnicas de escala, torneio, modelos elitistas e seleção por classificação (GOLDBERG, 1989; MICHALEWICZ, 1994). Neste trabalho, assim como Liao, *et al*., 2006 e Tseng, *et al*., 2009, usou-se a seleção roleta para eleger soluções potencialmente úteis para a próxima geração. Baseando-se no nível de aptidão, a *roulette wheel selection* associa uma probabilidade de seleção a cada cromossomo individual da população, conforme pode ser visualizado na equação (15) (HOLLAND, 1975).

$$P\,[\text{Indivíduo } i \text{ é escolhido}] = \frac{F_i}{\sum_{j=1}^{G} F_j} \qquad (15)$$

onde F_i é igual à aptidão do indivíduo i e G é o tamanho da população.

A *roulette wheel selection* padrão pode ser imaginada como uma roleta de um cassino na qual uma proporção da roda (abertura de roleta) é atribuída a cada um dos indivíduos, dimensionada de acordo à sua adequação. Para isso, basta dividir a aptidão do indivíduo pela aptidão total da população, normalizando-as para 1. Assim, uma seleção aleatória é feita de forma semelhante ao modo como a roleta é girada. Portanto, cromossomos com maior valor de adequação têm uma maior probabilidade de contribuir com a próxima geração (LIAO *et al*., 2006).

2.4.3.2 Crossover

O segundo operador (*crossover*) tem a função de mesclar as características de um par de cromossomos (progenitores) para gerar um novo par de descendentes (filhos) e desta forma, permite uma diversificação no espaço de soluções ao gerar configurações diferentes (BARAGONA, 2001). Para realizar essa operação, geralmente, os pares de cromossomos progenitores são escolhidos aleatoriamente com determinada probabilidade (LIAO *et al*., 2006).

Vamos considerar, a título de exemplo, o seguinte cromossomo binário 1101001100101101. A cadeia representaria uma possível solução para algum problema de otimização de parâmetros. Novos pontos de amostra no espaço são gerados pela recombinação de duas cadeias progenitoras. Considere 0001110110010000 como sendo o segundo progenitor. Usando um único ponto de recombinação escolhido aleatoriamente, o *crossover* ocorre da seguinte maneira:

11010 | 01100101101 (progenitor 1 com um ponto de corte)

00011 | 10110010000 (progenitor 2 com um ponto de corte)

Trocando os fragmentos entre os dois progenitores, a seguinte prole é gerada:

1101010110010000 (filho 1)

0001101100101101 (filho 2)

2.4.3.3 Mutação

Após o *crossover*, aplica-se o operador de mutação. Este último operador altera um ou mais genes do cromossomo de maneira aleatória, geralmente com uma pequena

taxa de mutação, de forma a evitar a obtenção de mínimos locais (BARAGONA, 2001). Para exemplificar esse operador, vamos considerar um dos progenitores do exemplo da seção anterior. A mutação toma como entrada um cromossomo e o devolve com alteração. Neste exemplo, será escolhido aleatoriamente um bit para realizar o complemento.

11010011**0**101101 (cromossomo de entrada)

11010011**1**101101 (cromossomo de saída)

Uma vez aplicados os operadores genéticos, avalia-se o desempenho dos novos indivíduos (nova geração) formados após os processos de *crossover* e mutação. Selecionam-se os melhores indivíduos e este processo é repetido até se atingir um determinado critério. (CHIOU, LAN, 2001).

2.4.4 Agrupamento e Reconhecimento de Padrões com Algoritmo Genético

O agrupamento de dados baseado em algoritmos genéticos explora a capacidade de pesquisa desses algoritmos para procurar centros de grupo apropriados no espaço de busca (MAULIK, BANDYOPADHYAY, 2000). Para implementar um GA em uma aplicação qualquer é necessário escolher a forma de representação dos cromossomos. Dois formatos básicos de codificação têm sido comumente utilizados em algoritmos genéticos. Um deles é o esquema binário adotado por Liao *et al*. (2006), Hall *et al*. (1999), Chiou, Lan (2001) e o outro é a codificação via número real (TSENG *et al*., 2009; WIKAISUKSAKUL, 2014; MAULIK, BANDYOPADHYAY, 2000; BANDYOPADHYAY, MAULIK, 2002; LIAO, 2002).

Dentre as poucas abordagens sobre o uso de algoritmos genéticos no agrupamento de séries temporais encontram-se as abordagens propostas por Baragona, 2001, Liao *et al*., 2006 e Tseng *et al*., 2009. Por sua vez, cada um destes trabalhos propõe diferentes representações para os cromossomos. Na abordagem de Baragona, 2001 o cromossomo possui uma quantidade de genes igual ao número de objetos, onde cada gene recebe um número inteiro positivo que estabelece o grupo ao qual um dado objeto pertence em uma dada solução. Na abordagem de Liao *et al*., 2006 os cromossomos representam os centros dos agrupamentos. O comprimento do cromossomo é obtido pelo número pré-determinado de grupos juntamente com o

número de dígitos usados para representar cada centro (codificação binária). Tseng *et al*., 2009 armazenam em cada cromossomo o resultado de uma segmentação possível para uma dada série temporal e o objetivo consiste no agrupamento de segmentos em diferentes grupos. Todas essas abordagens utilizam alguma medida de distância como critério para avaliar a semelhança entre duas séries, sendo a distância Euclidiana o critério mais comumente usado pelos autores (TSENG *et al*., 2009).

A Tabela 1 traz mais informações sobre os três trabalhos citados. Em todos eles trata-se de séries temporais univariadas. Não são encontrados trabalhos de agrupamento e reconhecimento de padrões que utilizem os algoritmos genéticos em séries temporais multivariadas. Este trabalho propõe uma abordagem para agrupamento de séries temporais multivariadas com algoritmo genético baseado no índice SPCA (PCA *Similarity Factor*) como métrica de similaridade.

Tabela 1 – Resumo dos trabalhos de agrupamento de séries temporais baseados em GA

Características dos trabalhos	**Baragona (2001)**	**Liao, et al. (2006)**	**Tseng, et al. (2009)**
Natureza das séries	Univariadas	Univariadas	Univariadas
Codificação do cromossomo	Inteiro positivo que representa o grupo ao qual cada série pertence	Números binários que representam os centros dos grupos	Números reais que representam uma segmentação possível para uma dada série temporal
População Inicial	Aleatória de tamanho passado como parâmetro	O algoritmo tem duas fases. Na primeira, são formados os agrupamentos iniciais, a partir do algoritmo *Merge_Sets_Finding*, a serem utilizados na segunda fase pelo algoritmo genético	São gerados aleatoriamente segmentos de mesmo comprimento e, então, valores de variância são adicionados aos pontos de corte para movê-los para a esquerda ou para a direita de maneira a formar os cromossomos iniciais. São utilizados 80 cromossomos
Função de avaliação	A dissimilaridade das séries foi	As séries são classificadas com	A dissimilaridade das séries é

	calculada a partir da correlação cruzada entre séries estimadas por modelos AR	base no centro mais próximo, baseando-se em duas medidas de distâncias: Euclidiana e *Dynamic Time Warping*	verificada por meio da distância Euclidiana. A transformada *wavelet* também é usada para ajustar o comprimento das subsequências, uma vez que estes comprimentos podem ser diferentes.
Seleção	Seleção aleatória com probabilidade proporcional a aptidão das soluções, aliada a estratégia elitista	Seleção Roleta Padrão	Seleção Roleta Padrão
Crossover	Parcialmente combinado com uma probabilidade de 0,6.	Ponto único de corte de acordo com uma probabilidade adaptativa	Ponto único de corte com taxa de 0,8
Mutação	Os genes são alterados de acordo com a probabilidade de mutação de 0,001.	Alteração de alguns bits do cromossomo de '0' para '1' ou '1' para '0' de acordo com a probabilidade adaptativa	Mutação em um único gene com taxa de 0,3
Término	Número pré-especificado de iterações	Critério não informado	Não deixa claro o critério de término, mas informa que o número máximo de iterações é 500

CAPÍTULO 3
METODOLOGIA

3.1 APLICAÇÃO DO ALGORITMO GENÉTICO PARA O AGRUPAMENTO E RECONHECIMENTO DE PADRÕES EM SÉRIES MULTIVARIADAS

A metodologia de agrupamento genético de séries temporais multivariadas desenvolvida neste trabalho foi aplicada a séries que representam processos industriais em dois estudos de caso. Em ambos, o objetivo é identificar operações normais e operações com falha. O primeiro estudo de caso trata-se de um processo industrial real e o segundo de um processo simulado. Os detalhes de cada estudo de caso serão abordados no próximo capítulo.

A metodologia compreendeu a codificação dos cromossomos por centros (números reais) e não pela matriz de pertinência. Essa escolha deu-se em virtude da natureza das séries multivariadas. Essas séries possuem um volume maior de dados a serem manipulados. Nesse contexto, codificar os cromossomos pela matriz de pertinência implica em, a cada iteração, não apenas atualizar a matriz de pertinência, mas também em calcular os centros de cada grupo para cada indivíduo da população. Por sua vez, na codificação pelos centros dos grupos, é necessário apenas calcular a matriz de pertinência, uma vez que os centros já são os próprios cromossomos.

Cada cromossomo (indivíduo) é formado por uma coleção de genes que constituem os centros dos grupos obtidos. Por sua vez, cada centro é uma série multivariada representada por uma matriz (eq. 8) (e não simplesmente um vetor no espaço multidimensional). O comprimento (ou tamanho) de cada cromossomo é determinado pela quantidade pré-especificada de grupos juntamente com a quantidade de dígitos usados para representar cada grupo. Consideremos como exemplo uma amostra de n objetos (n séries multivariadas), cada objeto composto por três séries

temporais associadas a 3 variáveis de processo e cada série temporal com oito instantes de medição (comprimento da janela de tempo). Deseja-se segregar as séries em dois grupos (normal e com falha). Cada cromossomo (ou candidato a solução) é composto por duas séries temporais multivariadas que representam os centros de cada um dos grupos. Dessa forma, como temos 8 instantes de medição e 3 variáveis para cada centro, cada cromossomo é composto por 48 genes (matriz 8x6). A Figura 7 ilustra o cromossomo.

Centro do grupo 1			Centro do grupo 2		
Var 1	Var 2	Var 3	Var 1	Var 2	Var 3
0,0231	0,3567	0,3020	-0,0057	0,2268	0,0707
0,0223	0,3618	0,2964	-0,0053	0,2298	0,0689
0,0285	0,3687	0,2907	0,0220	0,2393	0,0691
0,0355	0,3773	0,2878	0,0469	0,2694	0,0842
0,0451	0,3835	0,2892	0,0998	0,2921	0,1076
0,0578	0,3882	0,2951	0,1147	0,3258	0,1596
0,0625	0,3932	0,3043	0,1125	0,3399	0,2114
0,0816	0,4010	0,3175	0,1221	0,3480	0,2567

Figura 7 – Representação de um cromossomo.

Na Figura 7, as linhas da matriz representam os instantes de medição (dentro da janela de tempo de cada série multivariada), as colunas de 1 a 3 (azuis) representam o centro do primeiro grupo (com as respectivas séries associadas a cada uma das variáveis) e as colunas de 4 a 6 (verdes) representam o centro do segundo grupo. Vale ressaltar que este trabalho traz a importante contribuição da estrutura de duas dimensões para o cromossomo, visto que não encontramos esse tipo de estrutura na literatura.

O algoritmo inicia gerando aleatoriamente a população inicial com uma quantidade definida de cromossomos (primeira geração). Cada geração compreende candidatos à solução do problema. Todas as gerações possuem a mesma quantidade de indivíduos/cromossomos que é o tamanho da população (G). A Figura 8 ilustra cada iteração do algoritmo genético.

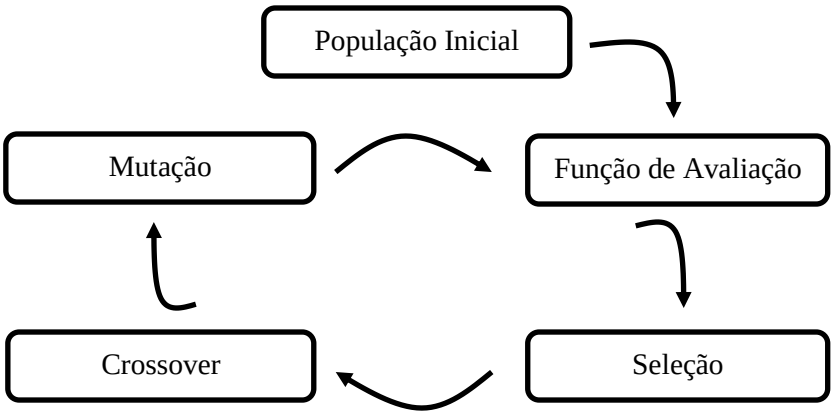


Figura 8 – Operações em cada iteração do GA (RIBEIRO e FONTES, 2017).

Cada indivíduo da população é submetido à função de avaliação. Quanto melhor for a solução para resolver o problema em questão, maior será o seu valor de aptidão. Em cada iteração todos os objetos da amostra são classificados ou agrupados para cada cromossomo (ou possível solução) de acordo com o centro mais próximo (menor distância conforme uma métrica de similaridade). As pertinências dos objetos a cada um dos centros presentes em cada cromossomo são obtidas através da eq. 5 (matriz de partição para o FCM probabilístico). Ou seja, cada cromossomo representa um conjunto de centros e, consequentemente, um resultado de agrupamento.

No primeiro estudo de caso a métrica de similaridade utilizada foi o índice SPCA modificado (Eq. 14). A função de avaliação de cada cromossomo foi baseada na função objetivo do algoritmo clássico FCM (Eq. 1), com uma extensão (ou restrição suave) para maximizar a distância dentre os centros/padrões dos respectivos grupos (Eq. 16) (centros do próprio cromossomo).

$$J(h_z = [v_1^z, \dots, v_1^z]) = \sum_{i=1}^{C}\sum_{k=1}^{N} u_{ikz}^m \cdot \left\|x_k\text{-}v_i^z\right\|^2 + \frac{1}{\sum_{i=1}^{C}\sum_{\substack{j=1 \\ j>i}}^{C}\left\|v_i^z - v_j^z\right\|^2} \qquad (16)$$

h_z $(h_z = [v_1^z, \dots, v_C^z])$ é o indivíduo z em uma dada geração $(z = 1, \dots, G)$, o que corresponde a um candidato à solução do problema (conjunto de centros/padrões que possuem uma aptidão estabelecida pela eq. 16). v_i^z $(i = 1, \dots, C)$ é o centro ou padrão do grupo i associado ao indivíduo z. C é o número de grupos, N é o número de objetos, m é o expoente de fuzzificação que controla o grau de sobreposição *fuzzy* $(m > 1)$, x_k é o k-ésimo objeto e μ_{ikz} é o grau de pertinência do objeto x_k ao centro do i-ésimo grupo

do indivíduo *z* (o que implica que cada indivíduo está relacionado a uma determinada matriz de pertinência). O objetivo da otimização consiste em minimizar o valor de *J* na eq. (16). A primeira parcela da eq. (16), da mesma forma que no FCM clássico, minimiza as distâncias entre elementos de um mesmo grupo (distâncias intragrupo) na tentativa de reconhecer grupos homogêneos. A segunda parcela representa uma restrição suave de forma a maximizar as distâncias entre os centros dos grupos (distância extragrupos), evitando a obtenção de indivíduos com fusão de grupos ou com uma elevada similaridade entre os seus respectivos centros/padrões.

No segundo estudo de caso, baseado no trabalho de Fontes e Budman (2017), aplicou-se uma métrica de similaridade híbrida que compreende o SPCA juntamente com outra métrica que considera o valor médio das séries que compõem o objeto. Essa nova métrica é denominada Distância Euclidiana Baseada na Média (*Average-based Euclidean Distance* – AED). A AED consiste em gerar um vetor para cada objeto (cada série multivariada) que terá como componentes as médias aritméticas de cada uma das séries que compõem o objeto. A AED entre dois objetos compreende a distância Euclidiana entre os seus respectivos vetores de média. Desta forma, a função objetivo (de onde se obtêm o *fitness* de cada cromossomo) ficou:

$$\min_{(\boldsymbol{U},\boldsymbol{V})} J_{\varepsilon}(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n}\left\{\alpha \cdot u_{ik}^{\varepsilon}\left\|\boldsymbol{X}_k - \boldsymbol{V}_i\right\|_{SPCA}^2 + (1-\alpha)\cdot u_{ik}^{\varepsilon}\left\|\overline{\boldsymbol{X}_k} - \overline{\boldsymbol{V}_i}\right\|_{AED}^2\right\} + \beta \cdot \sum_{i=1}^{c}\sum_{\substack{j=1 \\ j>i}}^{c}\frac{1}{\left\|V_j - \boldsymbol{V}_i\right\|_{HYB}^2} \quad (17)$$

onde

$$\left\|V_j - \boldsymbol{V}_i\right\|_{HYB} = \alpha \cdot \left\|V_j - \boldsymbol{V}_i\right\|_{SPCA} + (1-\alpha)\cdot\left\|\overline{V_j} - \overline{\boldsymbol{V}_i}\right\|_{AED}$$

Sujeito a:

$$\boldsymbol{U} \in \Re^{c\times n}:\ u_{ik} \in [0,1],\ \sum_{k=1}^{n} u_{ik} > 0\ \ \forall i \quad \text{and} \quad \sum_{i=1}^{c} u_{ik} = 1\ \ \forall k \quad (18)$$

onde $\boldsymbol{X}_k$ (k=1,…,n) é um objeto (MTS), $\boldsymbol{V}_i$ (i=1,…, c) é o centro de um grupo $\left(\boldsymbol{X}_k \text{ and } \boldsymbol{V}_i \in \Re^{m\times p}\right)$ e $\boldsymbol{V}$ é o conjunto de matrizes dos grupos $\{\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_c\} \in \Re^{m\times p}$. ($m$ é o número de medições e p é o número de variáveis) $\left\|\boldsymbol{X}_k - \boldsymbol{V}_i\right\|_{SPCA}$ é a distância entre o objeto X_k e o centro do grupo i com base na métrica SPCA (eq. 14). $\overline{X_k}$ e $\overline{V_i} \in \Re^{p}$ são vetores compostos das médias associadas a cada elemento (cada série temporal) do objeto e do centro do grupo, respectivamente. $\left\|\overline{\boldsymbol{X}_k} - \overline{\boldsymbol{V}_i}\right\|_{AED}$ é a distância Euclidiana baseada na média (AED) entre esses vetores. α $(\alpha \in [0,1])$ é um parâmetro de sintonia

(*tradeoff parameter*). β é um outro parâmetro de sintonia que pondera a separação entre os centros dos diferentes grupos. Ambos α e β foram determinados através de validação cruzada com base nos resultados de classificação. Os melhores resultados foram obtidos com $\alpha=0,5$ e $\beta=1.10^{-4}$em ambas as abordagens (FCM-GA e FCM com otimização clássica).

A função de avaliação (função objetivo) (eq. 17) continua adotando a distância entre os centros (distância extragrupos). A diferença na implementação do primeiro estudo de caso está na distância usada para avaliar a similaridade entre cada objeto e o centro de cada grupo. Essa nova implementação se fez necessária na medida em que a métrica SPCA de forma isolada não foi capaz de reconhecer a diferença entre os objetos de falha e normais (FONTES e BUDMAN, 2018; FONTES e BUDMAN, 2017).

O FCM aplicado a séries temporais univariadas adota como métrica padrão de similaridade a distância Euclidiana. Esta métrica não foi utilizada neste trabalho uma vez que as séries são multivariadas e, além disso, possuem tamanhos diferentes (os objetos estão associados a diferentes comprimentos de janela de tempo).

A função de avaliação classifica as diferentes soluções/indivíduos da população através de um critério de qualidade (valor de aptidão) ordenando os indivíduos conforme sua aptidão. Para o primeiro estudo de caso, a aptidão de cada indivíduo (candidato aos centros ou solução do problema) é simplesmente o inverso do valor da função obtido pela eq. 16:

$$A(h_z) = \frac{1}{J(h_z)} \tag{19}$$

$A(h_z)$ é a aptidão do indivíduo h_z. Para o segundo estudo de caso vale a mesma regra, entretanto com a eq. 17.

Em seguida é aplicado o operador de seleção. Inspirado na evolução natural, este operador seleciona as melhores soluções da geração atual de forma a criar uma nova geração mais apta, perpetuando as "espécies" com os melhores genes. Dessa maneira, consegue-se eliminar as soluções menos aptas da população. Assim como Baragona (2001), a estratégia de seleção adotada neste trabalho foi do tipo roleta com abordagem elitista. Nesse tipo de seleção todos os cromossomos da população (6 neste exemplo) são colocados em uma roleta e cada um ocupa uma secção proporcional ao respectivo valor de aptidão dado pela eq. 19 (Figura 9).

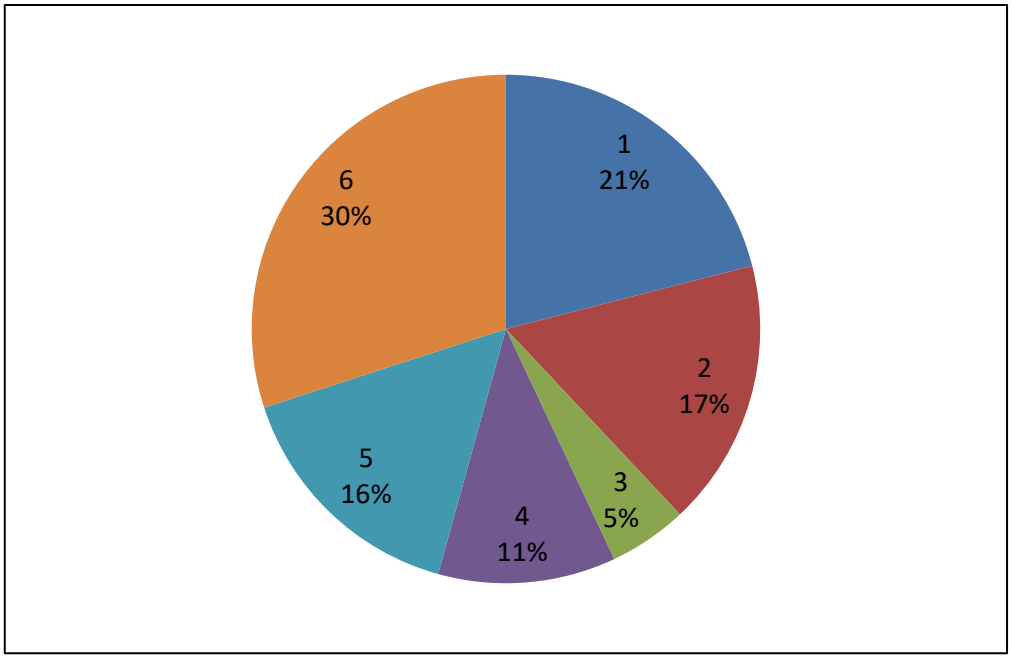


Figura 9 – Técnica do tipo roleta para seleção de cromossomos.

A roleta é então "girada" e, evidentemente, os cromossomos com maior aptidão obtêm maior probabilidade de serem selecionados e, assim, propagados para a geração seguinte. A nova geração sempre tem a mesma quantidade de indivíduos da atual. Este procedimento não impede que o mesmo indivíduo seja selecionado repetidas vezes.

A implementação da seleção roleta foi realizada por meio da função *roulette wheel selection* disponibilizada no site oficial do *Matlab* (ABOUSLEIMAN, 2015). Esta função recebe como entrada um vetor de probabilidades proporcional aos valores de aptidão dos indivíduos da população (saída da função de avaliação). Considere, por exemplo, seis cromossomos com suas probabilidades armazenadas em um vetor $V = \{0.210, 0.170, 0.050, 0.113, 0.157, 0.300\}$.

O algoritmo basicamente faz com que cada cromossomo tenha uma quantidade de posições na roleta proporcional à sua probabilidade. Para obter a quantidade total de posições da roleta (n) tem-se:

$$n = \frac{1}{min(V)} \tag{20}$$

Utilizando a eq. 20 nesse exemplo o valor obtido para n é *20*. A Figura 10 representa a roleta com as 20 posições. Caso o valor obtido para n não seja inteiro, o valor é arredondado para o inteiro menor mais próximo.

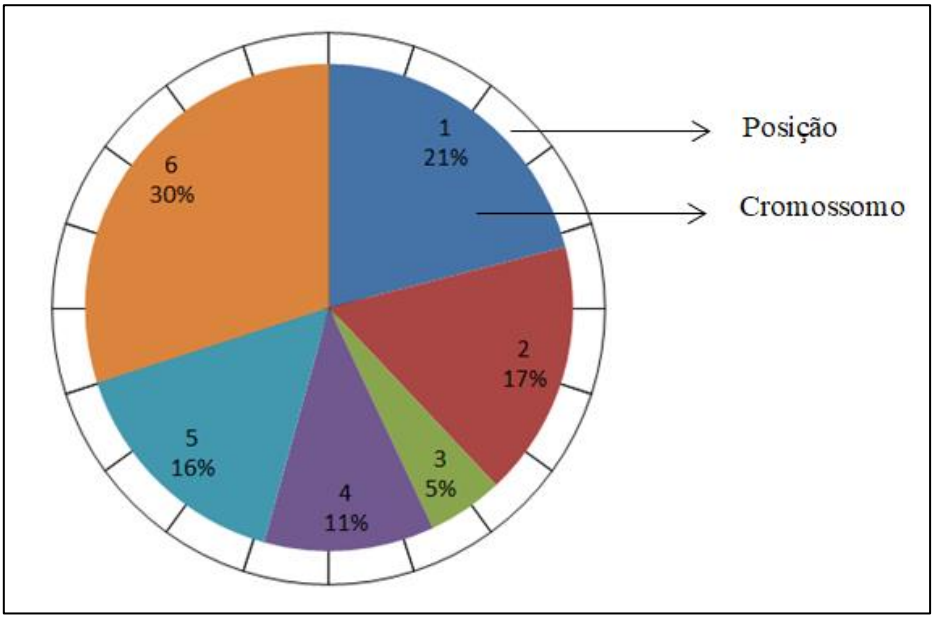


Figura 10 – Roleta com 20 posições.

Na Figura 10 o cromossomo 6 ocupa a maior quantidade de posições (seis), uma vez que possui a maior probabilidade. A eq. 21 é usada para determinar o vetor com as quantidades de posições referentes a cada cromossomo:

$$V' = \frac{1}{\min(V)} * V \tag{21}$$

Um número inteiro aleatório que representa a posição da roleta (valor entre 1 e 20, no exemplo) é gerado e o cromossomo cujo segmento abrange essa posição é selecionado. Por exemplo, se o valor sorteado for 13, podemos ver na roleta que o número 13 é terreno do cromossomo 5. O cromossomo 5 é a saída da função *roulette wheel selection*.

A cada iteração essa função é chamada um número de vezes igual ao tamanho da população de forma a manter a mesma quantidade de indivíduos a cada geração. Apesar da seleção roleta elevar a probabilidade de que a melhor solução seja selecionada para a próxima geração, não há garantia de que isso aconteça. Por esse motivo, além da seleção roleta, a estratégia elitista foi adotada de maneira a preservar as melhores soluções da população. O elitismo garante que a melhor solução se manterá na próxima geração. O melhor cromossomo (a melhor solução) é copiado para a nova população/geração e os demais indivíduos são selecionados através da roleta.

Em seguida o operador de *crossover* é aplicado. Este operador, juntamente com a mutação, viabiliza a melhor exploração da região de busca e um melhor desempenho do método do GA que pode ser materializado através da obtenção de um melhor mínimo local (melhor solução para o problema).

No *crossover* são escolhidos aleatoriamente pares de cromossomos (indivíduos) diferentes a serem cruzados. A finalidade é combinar características de duas soluções (progenitores) de maneira a gerar duas novas soluções (filhos) que herdem essas características. Esse operador é aplicado com probabilidade dada pela taxa de *crossover* (Pc). Para realizar esta operação são definidos os pontos de corte para os cromossomos de maneira aleatória. Como cada cromossomo é composto pelas séries temporais multivariadas que representam o centro de cada grupo, se tivéssemos, por exemplo, dois grupos, então existiriam dois pontos de corte (p_1 e p_2). Neste exemplo, o primeiro filho (Filho 1) é gerado em duas etapas, visto que temos dois centros. A primeira etapa compreende a definição do centro do primeiro grupo. Ela é realizada copiando os genes do primeiro progenitor (Progenitor 1) desde o começo do cromossomo até p_1 e o restante é copiado do segundo progenitor (Progenitor 2). A segunda etapa, ainda para a geração do Filho 1, compreende a definição do centro/padrão associado ao segundo grupo, sendo realizada de modo análogo à primeira, porém com outro ponto de corte p_2. Conforme pode ser visualizado na Figura 11, a ordem é trocada para o segundo filho (Filho 2).

O procedimento de *crossover* estabelecido possui duas características importantes:

a) Permite apenas o cruzamento de genes ou partes das séries temporais associadas às mesmas variáveis. Não há a possibilidade de que perfis dinâmicos relacionados a diferentes variáveis de processo sejam permutados. Ou seja, uma série de vazão de gás do Progenitor 1 não pode ser cruzada com uma série da temperatura de entrada do Progenitor 2, por exemplo;
b) As séries permutadas referem-se ao centro do mesmo grupo.

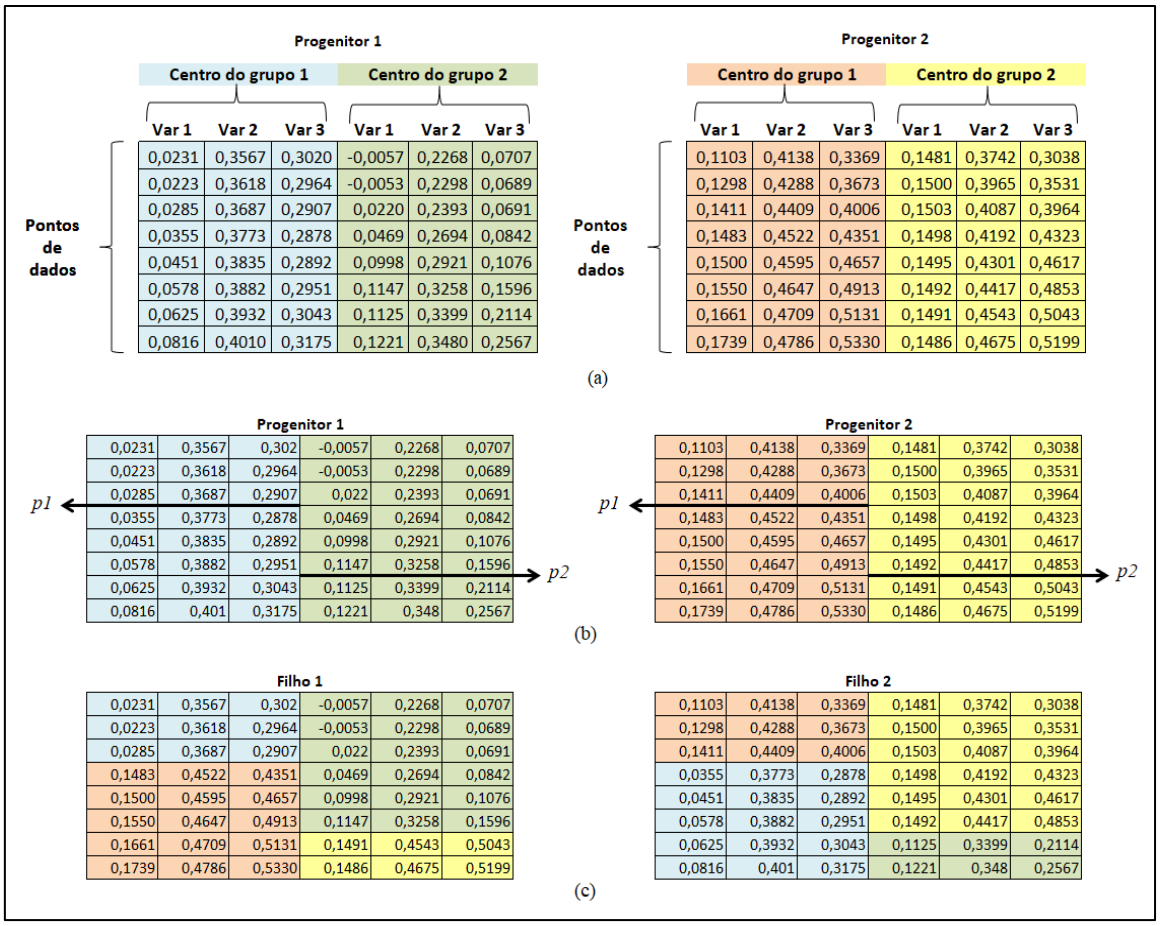


Figura 11 – Crossover.

(a) dois indivíduos são escolhidos. (b) um ponto de corte para cada grupo é escolhido. (c) são recombinadas as características, gerando dois novos indivíduos.

No passo final do *crossover* os filhos substituem seus progenitores na população, alterando-a. Em seguida, aplica-se nesta nova população o último operador do algoritmo genético, o operador de mutação. Este operador altera aleatoriamente alguns genes do cromossomo de acordo com a taxa de mutação (P_m). Em outras palavras, P_m é a probabilidade de mutação dos cromossomos da população. A Figura 12 ilustra a aplicação desse operador.

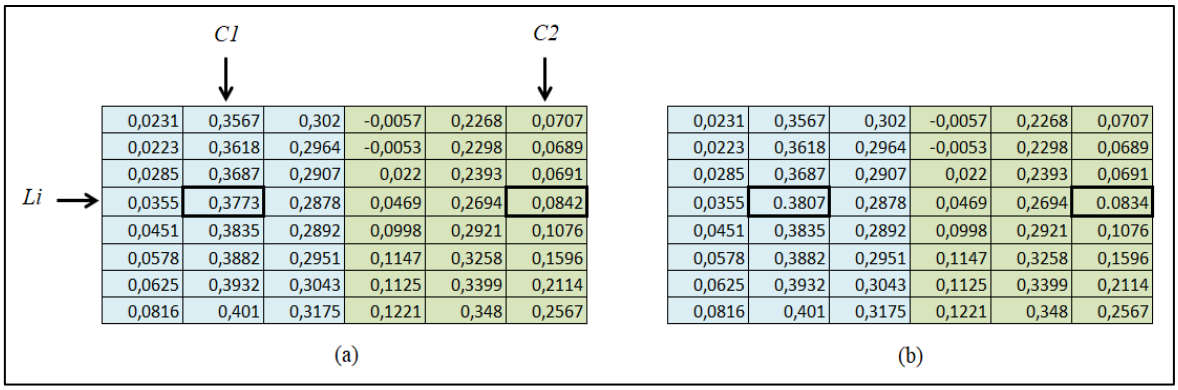


Figura 12 – Mutação.

(a) um gene (um ponto de uma série) aleatório para cada centro é escolhido. (b) os genes são alterados, obtendo-se um novo indivíduo.

Aplica-se a mutação a tantos genes quanto forem a quantidade de centros/grupos considerados. Continuando com o exemplo que possui dois centros, são alterados dois genes. Para escolha dos genes a serem mutados (Figura 12), uma linha (*Li*) é aleatoriamente escolhida. Em seguida, escolhe-se, também de maneira aleatória, uma coluna (*C1*) para o centro 1 e uma coluna (*C2*) para o centro 2. Então, os genes referentes a linha Li e as colunas *C1* e *C2* são alterados, conforme visualizado na Figura 12. Após a mutação, a função de avaliação é novamente aplicada sobre toda a nova geração e a próxima iteração é iniciada com uma nova população selecionada.

As possíveis soluções se movem de geração em geração e o algoritmo finaliza quando a solução (melhor indivíduo da população) mantiver seu nível de aptidão sem alterações. O resultado final do algoritmo fornecerá os centros, ou padrões, dos grupos formados, com os quais é possível obter a matriz de pertinência final (graus de pertinência de todos os objetos) através da eq. (5). A Tabela 2 apresenta os valores dos parâmetros adotados no algoritmo genético.

Tabela 2 – Valores dos parâmetros do algoritmo genético

Parâmetro	**Valor**
Tamanho da população	6
Crossover	Taxa de 0,6
Mutação	Taxa de 0,02

Como pode ser visto na Tabela 2, o tamanho adotado para a população foi 6 ($G = 6$). Testes com valores superiores a 6 foram realizados e os resultados mostraram

que o GA convergia para o mesmo resultado, entretanto as iterações demoravam mais tempo para serem concluídas. A taxa de 0,6 para o crossover significa que 60% dos cromossomos da população são escolhidos de maneira aleatória para o *crossover*. Essa taxa foi escolhida, em virtude de ser um percentual que está dentro das taxas utilizadas nos poucos trabalhos da literatura que abordam o uso do GA para o agrupamento de séries temporais (BARAGONA, 2001; TSENG, *et al*., 2009). A taxa de mutação de 0,02 foi escolhida de tal forma que houvesse um equilíbrio entre as características da nova geração e as características de seus progenitores provocando também uma alteração na geração atual, visando uma melhor exploração da região de busca.

CAPÍTULO 4

ESTUDOS DE CASO E RESULTADOS

4.1 ESTUDO DE CASO 1: DETECÇÃO DE FALHAS EM UMA TURBINA A GÁS

4.1.1 Descrição

O primeiro estudo de caso compreendeu a análise e detecção de falhas em uma turbina a gás existente na Unidade Termelétrica Rômulo Almeida (UTE-RA), integrante do parque industrial da Petrobras. Esta Unidade (Figura 13) está localizada no município de Camaçari-BA e possui três caldeiras de recuperação (*Heat Recovery Steam Generator* - HRSG), uma caldeira auxiliar (CA), uma turbina a vapor (TV), uma torre de resfriamento e três turbinas a gás (TG).

Figura 13 – Termelétrica Rômulo Almeida (SÁ BARRETTO, 2009).

Uma UTE é uma instalação industrial usada para geração de energia elétrica que é produzida a partir de insumos como gás combustível, óleo, carvão, vapor, entre outros. A UTE-RA tem como principal insumo o gás natural, sendo 137 MW a sua capacidade máxima de geração de energia elétrica. Desse total, 27MW é a capacidade de geração de cada TG e 56MW é a capacidade máxima de geração da TV. A Unidade opera em ciclo combinado e produz ainda 260.3 t/h de vapores de baixa e alta pressão.

Na Unidade Rômulo Almeida as turbinas a gás representam o principal componente do processo uma vez que são responsáveis pela maior quantidade de energia elétrica gerada (81 MW). As turbinas são produzidas pela Rolls-Royce (modelo RB211-G62 DF) (Figura 14).

Figura 14 – Turbina a gás RB211-G62 (ROLLS-ROYCE, 2010).

Uma turbina a gás é uma turbomáquina, ou seja, trata-se de uma máquina onde a energia é fornecida ou extraída por meio de um eixo giratório. Os insumos da turbina são o ar e um combustível. O processo fundamental da turbina está baseado na conversão de energia térmica em energia mecânica e desta em energia elétrica. A produção da energia térmica acontece por meio do processo da combustão que ocorre sobre a mistura de ar e combustível (gás natural). Dessa maneira, o fluido de trabalho será o conjunto de produtos da combustão. A conversão da energia mecânica em elétrica ocorre através do acionamento mecânico de um gerador elétrico acoplado ao eixo da turbina (SARAVANAMUTTOO *et al*., 1996).

Cada turbina é composta por três equipamentos: compressor, câmara de combustão e expansor. Inicialmente, o ar é admitido pelo compressor, onde ocorre a compressão. Comprimido, o ar é direcionado para a câmera de combustão onde se mistura com o gás natural. Uma chama piloto é utilizada para iniciar a combustão. Os gases produzidos na saída da câmara, à alta pressão e temperatura, se dilatam conforme passam pelo expansor da turbina que é acoplado a um gerador de energia, convertendo

energia mecânica em energia elétrica. Trata-se de um ciclo termodinâmico aberto, isto é, o ar é admitido na pressão atmosférica e os gases de escape após passarem pela turbina são descarregados na atmosfera sem que sejam realimentados no equipamento.

Os vapores de alta e baixa pressão produzidos pela UTE-RA são fornecidos a diversas indústrias do Complexo Petroquímico de Camaçari. A energia elétrica produzida é distribuída para o consumo interno, demandas de clientes externos e exportação para o Sistema Interligado Nacional (SIN). No SIN o Operador Nacional do Sistema Elétrico (ONS) coordena e controla a operação e transmissão de energia elétrica. É acordado com o ONS, que se houver qualquer ocorrência que impeça o fornecimento de energia por parte da UTE, a unidade estará sujeita a aplicação de multas contratuais. Este cenário determina que haja um controle sobre o processo de forma que seja possível prever prováveis falhas que ocasionariam suspensão do processo de fornecimento. Uma parada operacional na Unidade significa prejuízo financeiro, além de outros riscos tais como danos aos equipamentos e acidentes.

Cada turbina a gás possui 17 sensores de temperatura radialmente dispostos ao longo da câmara de combustão. Existe ainda um sistema de controle projetado para desarmar o equipamento (parada da turbina) caso algum destes sensores indique um valor de temperatura de, no mínimo, 150º C acima ou abaixo da média entre as temperaturas dos demais sensores. O objetivo é proteger o equipamento de possíveis danos causados por dilatação diferencial. Este evento caracteriza um *trip* (falha) por dispersão de temperatura (BARRAGAN, 2016) e não há, até o momento, uma explicação ou entendimento claro a respeito da fenomenologia desta ocorrência.

A detecção da falha na prática é realizada tardiamente, ou seja, a falha somente é verificada após o desarme automático do equipamento. Não há qualquer informação prévia sobre o padrão da falha ou mesmo uma definição sobre as variáveis úteis na predição da falha (FONTES e PEREIRA, 2016). Neste contexto, assumiu-se para este estudo de caso a aplicação de um algoritmo de agrupamento baseado em algoritmo genético e FCM. Apesar de existirem vários tipos de falha, este estudo de caso restringiu-se especificadamente à falha (*trip*) devido a dispersão de temperatura durante a partida da turbina.

4.1.2 Procedimento e coleta de dados

Etapa 1. Obtenção das amostras. Os dados foram obtidos a partir de um sistema de gerenciamento de informações disponível na UTE-RA, o PIMS (*Process Information Management System*). O banco de dados disponibilizado pelo PIMS incluiu o período de 2008 a 2010. Os dados estão organizados em 70 séries temporais multivariadas, que representam os eventos de partida da turbina. As três variáveis de processo consideradas para o estudo são a vazão de entrada de gás natural, a temperatura de entrada e a temperatura de saída do gás natural da turbina;

Etapa 2. Análise de similaridade intra e extragrupos. Esta etapa envolveu o cálculo do nível de similaridade/dissimilaridade entre as séries multivariadas (ou objetos), sendo o SPCA a métrica utilizada para esta análise;

Etapa 3. Agrupamento e reconhecimento de padrões de operação presente nos dados. Para realizar o agrupamento das séries temporais multivariadas foi utilizada a abordagem proposta baseada em algoritmo genético e *fuzzy c-means*. O método foi aplicado a um subconjunto (dados de treinamento) de objetos da amostra global;

Etapa 4. Classificação das séries. Após obtenção dos padrões e grupos presentes nos dados de treinamento, foi realizada a classificação das séries do conjunto de treinamento e teste (objetos não utilizados na etapa de agrupamento). O esquema abaixo (Figura 15) resume as etapas.

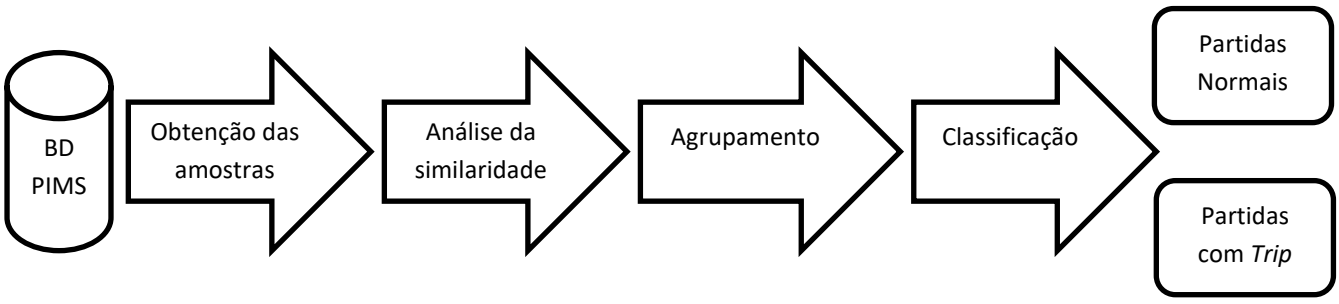


Figura 15 – Etapas da metodologia – caso 1.

4.1.3 Resultados e Discussão

Dentre as 70 séries temporais multivariadas (objetos) da amostra, apenas 10 referem-se a eventos de *trip* (falha) devido a dispersão de temperatura na partida da turbina e as demais (60) referem-se a eventos de operação normal de partida do equipamento. Tem-se, portanto, uma amostra desbalanceada com escassez de informação referente ao comportamento de falha, o que, por si só, representa mais um aspecto dificultador para este estudo de caso.

O comprimento das janelas de tempo associadas a cada objeto (cada série multivariada) não foi o mesmo para todos os objetos da amostra o que implica em um problema de agrupamento e classificação de séries temporais multivariadas com diferentes números de medições. Isto se deve ao fato de que as ocorrências de partida da turbina, verificadas ao longo de 2008 a 2010, não tiveram a mesma duração conforme efetivamente ocorre na prática quando se analisa o transitório de partida de equipamentos. A máxima duração (maior janela de tempo) entre os objetos coletados foi de 16 min e a menor foi de 6 min, todos com um período de amostragem de 0,5 min. Por sua vez, foi considerado o maior comprimento de janela de tempo (16 min) para representar cada série multivariada que define o centro ou padrão de cada grupo. A métrica de similaridade utilizada foi o SPCA, uma vez que esta consegue quantificar a similaridade entre séries com diferentes comprimentos através da comparação entre as direções de seus respectivos componentes principais.

A Figura 16 apresenta todas as séries da amostra de treinamento, com a ilustração dos perfis das 3 variáveis analisadas, quais sejam, vazão de alimentação de gás natural, temperatura de entrada e temperatura de saída do gás natural, respectivamente.

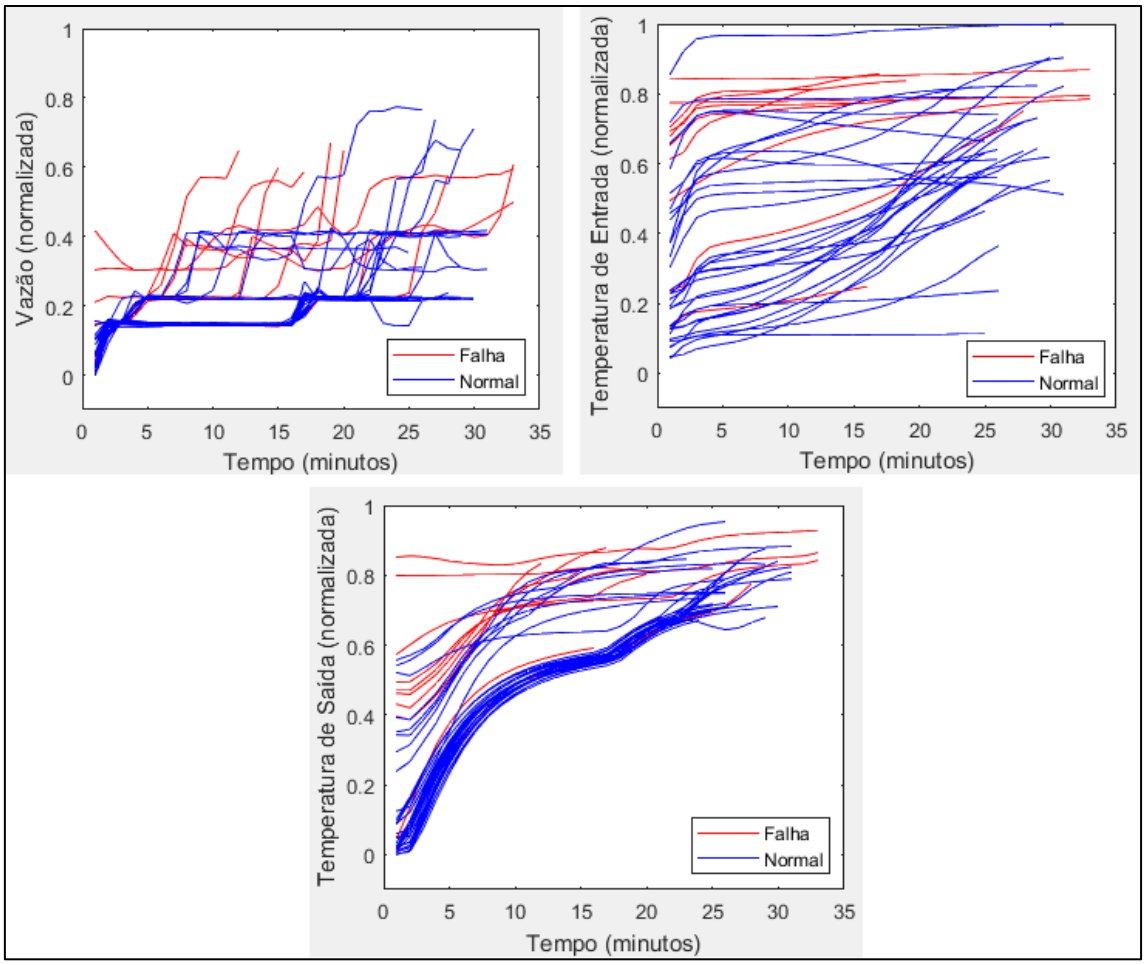


Figura 16 – Séries da amostra de treinamento.

Um objeto normal e um de falha estão apresentados nas Figuras 17 e 18. Cada série temporal da amostra foi normalizada considerando os valores máximos e mínimos de cada uma das variáveis presentes em toda a amostra (70 objetos). Ou seja, não foi efetuada normalização pelos valores máximo e mínimo de cada série individualmente, o que levaria a desconsiderar os diferentes níveis de operação praticados e modificaria o comportamento dinâmico original das variáveis. Vale salientar que a normalização dos dados possibilita a manutenção do sigilo dos dados da empresa.

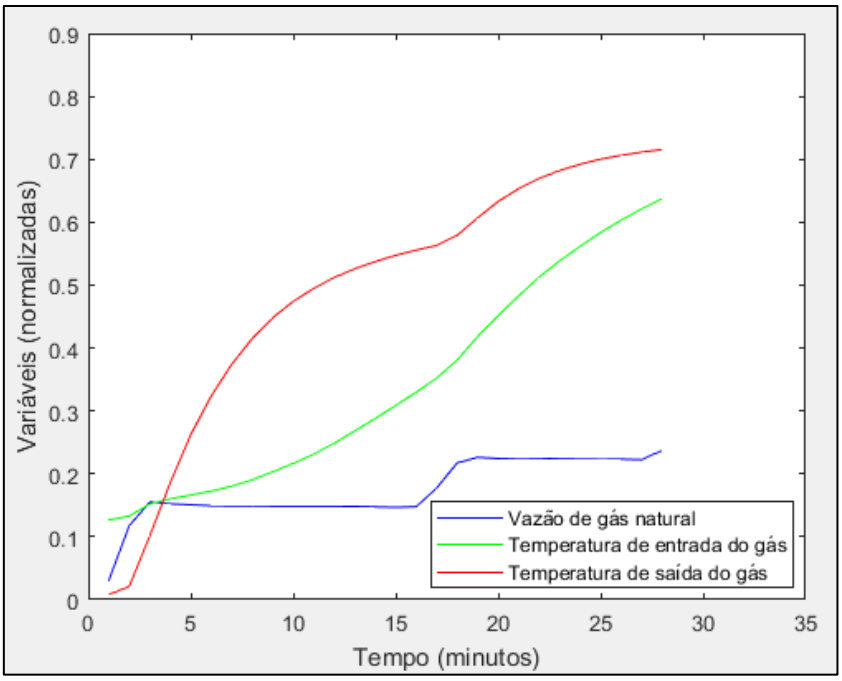


Figura 17 – Exemplo de partida normal.

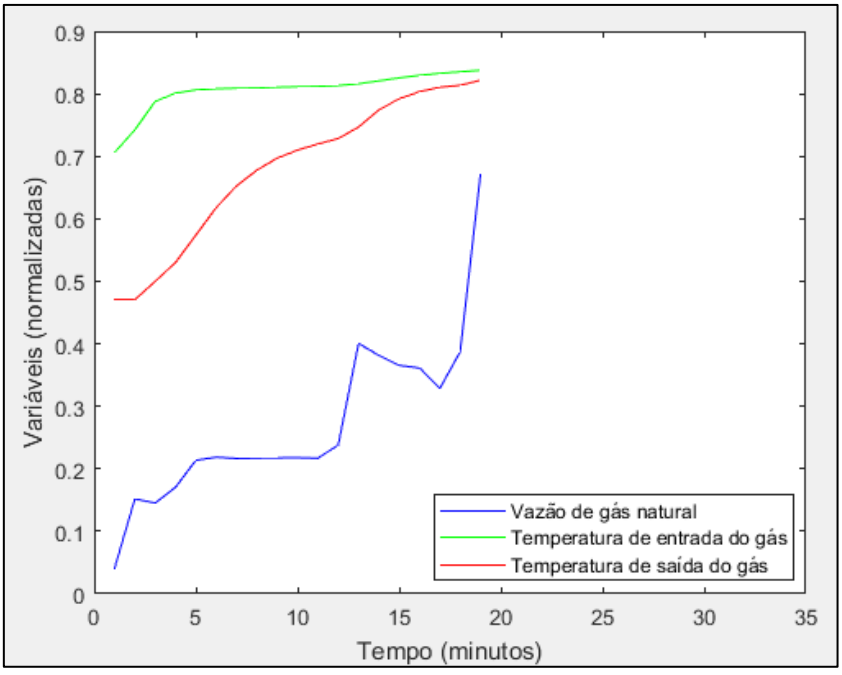


Figura 18 – Exemplo de partida com falha.

A partir das Figuras 16 a 18 é possível ter uma percepção da complexidade do problema. Nota-se a partir do conjunto de eventos que visualmente não é possível estabelecer uma distinção entre o comportamento dinâmico das variáveis associadas a eventos normais e a eventos de falha. Isso reforça a utilização de um algoritmo de agrupamento de objetos e de reconhecimento de padrões de falha e de normalidade que

possam representar uma referência e serem capazes de viabilizar uma tomada de decisão antecipada à falha do equipamento.

Ao aplicar o algoritmo genético a convergência é alcançada quando a melhor solução da população mantiver seu nível de aptidão constante por sucessivas iterações. O crescimento da aptidão do melhor indivíduo no decurso do processo evolutivo, bem como sua estabilização podem ser visualizados na Figura 19.

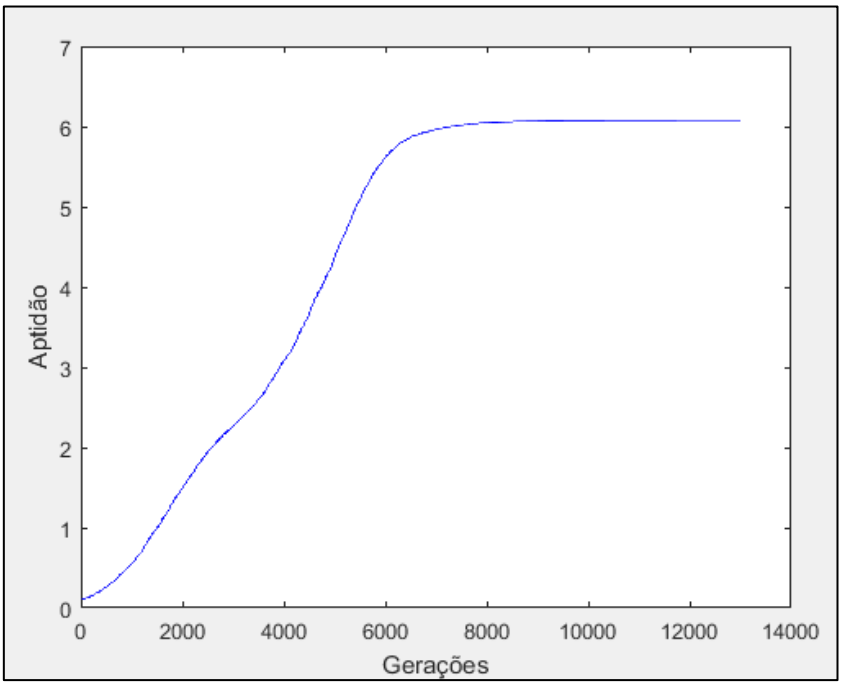


Figura 19 – Estabilização da aptidão da melhor solução ao longo das gerações.

Outra questão importante a ser analisada no algoritmo genético foi quanto ao uso do operador de mutação. Isso porque mesmo adotando-se uma taxa de mutação de 0,02, foi necessário limitar em ±1% a taxa de alteração de cada gene em relação ao valor atual para evitar a obtenção de centros com oscilações excessivas, tal como ilustrado na Figura 20.

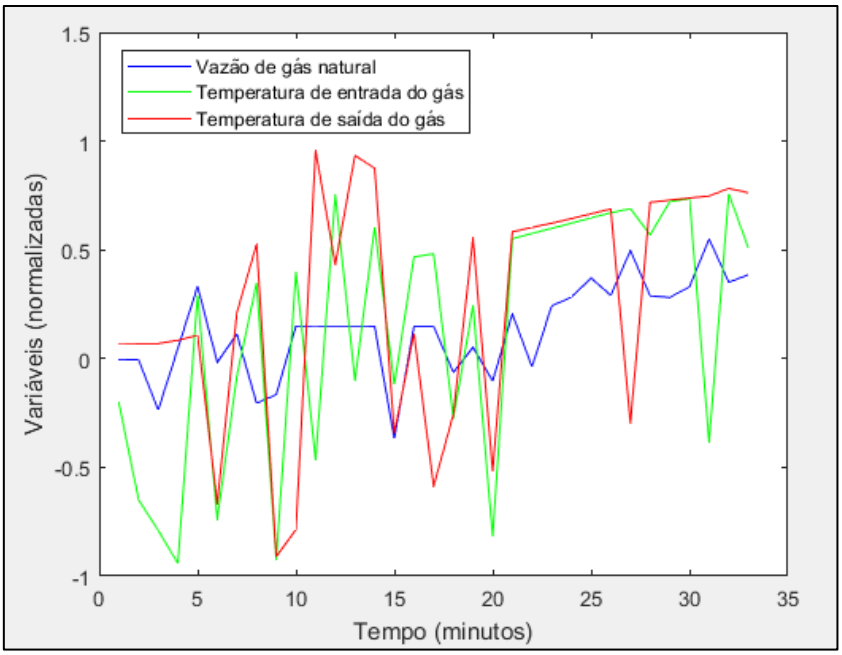


Figura 20 – Centro com oscilações excessivas (ausência de limitação para as alterações de genes através de mutação).

É importante destacar que a taxa de 1% não se refere à quantidade de genes que devem ser modificados, mas sim ao limite que é estabelecido na alteração em um dado gene quando este sofre mutação. Dessa forma, garante-se uma consistência nas soluções obtidas em relação ao comportamento dinâmico das variáveis e, ao mesmo tempo, evita-se que o GA fique estacionado em um mínimo local indesejável.

A amostra foi particionada em dados de treinamento e de teste. Os dados de treinamento são constituídos por 40 objetos, 30 de operação normal e 10 de operação com falha e os de teste, por sua vez, são compostos por 30 objetos, todos de operação normal (Tabela 3).

Tabela 3 – Amostras de treinamento e de teste

Amostra	**Total de Operação com Falha**	**Total de Operação Normal**
Original	10	60
Treinamento	10	30
Teste	0	30

Foram considerados três grupos ($C = 3$) nesse problema de otimização uma vez que esta partição ofereceu as menores taxas de erro de classificação. Com dois grupos as taxas de erro de classificação para os objetos normais foram elevadas o que sugere a

existência de um terceiro padrão de operação normal. Para $C > 3$ foram obtidos grupos com uma pequena quantidade de objetos, o que não sugere a existência de um número maior do que três grupos para classificar o fenômeno analisado.

Os grupos e centros foram obtidos usando a amostra de treinamento e, em seguida, os objetos de teste foram utilizados para validar os resultados de classificação. A Figura 21 apresenta o resultado final de classificação dos 40 objetos da amostra de treinamento, tendo-se três grupos com seus respectivos padrões (Figuras 22 a 24).

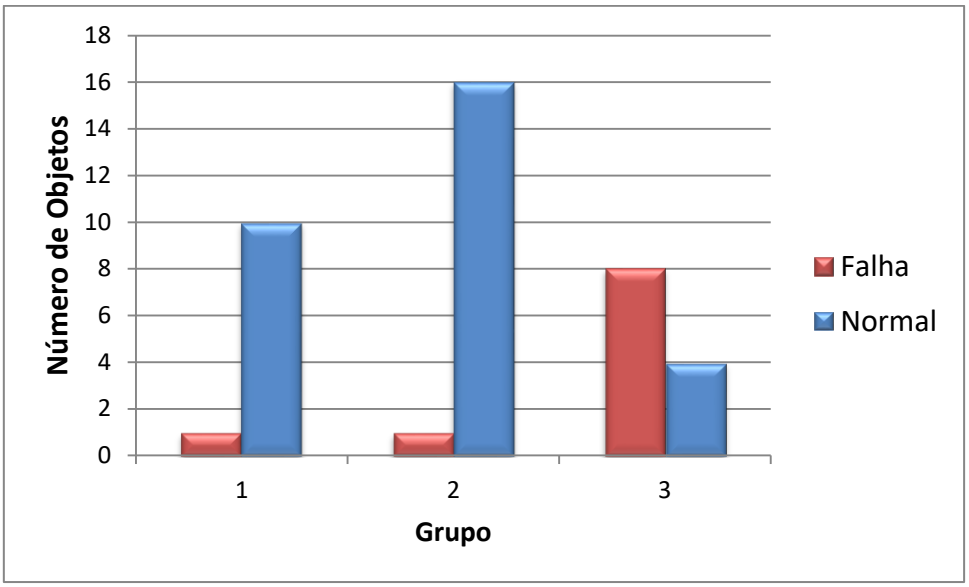


Figura 21 – Distribuição dos objetos nos três grupos.

A partir da Figura 21 pode-se notar a quantidade de objetos normais e com falha que foram alocadas em cada grupo, donde se verifica que nos grupos 1 e 2 a maioria dos objetos são normais, enquanto que o grupo 3 concentrou a maioria dos objetos de falha (80%). Isso sugere que o grupo 3 seja categorizado como tipicamente um grupo de falha. Por sua vez, os grupos 1 e 2 podem ser caracterizados como grupos de partida normal, o que evidencia inclusive a existência de mais de um padrão associado à operação normal.

As taxas de erro de classificação do FCM-GA, tanto na amostra de treinamento quanto na amostra de teste, para as operações normais e com falha estão apresentadas na Tabela 4.

Tabela 4 – Porcentagem de classificações erradas (FCM-GA)

Amostra	Operação com Falha	Operação Normal
Treinamento	20%	10%
Teste	—	3,3%

Os padrões reconhecidos (de falha e de operação normal) constituem um potencial para a utilização em sistemas inteligentes de controle ou de detecção e diagnóstico de falhas aptos a acompanhar, em tempo real, a probabilidade de falha do equipamento no decorrer de sua operação. Os três padrões obtidos pelo GA (ou centros dos agrupamentos), dois normais e um de falha, estão apresentados nas Figuras 22 a 24.

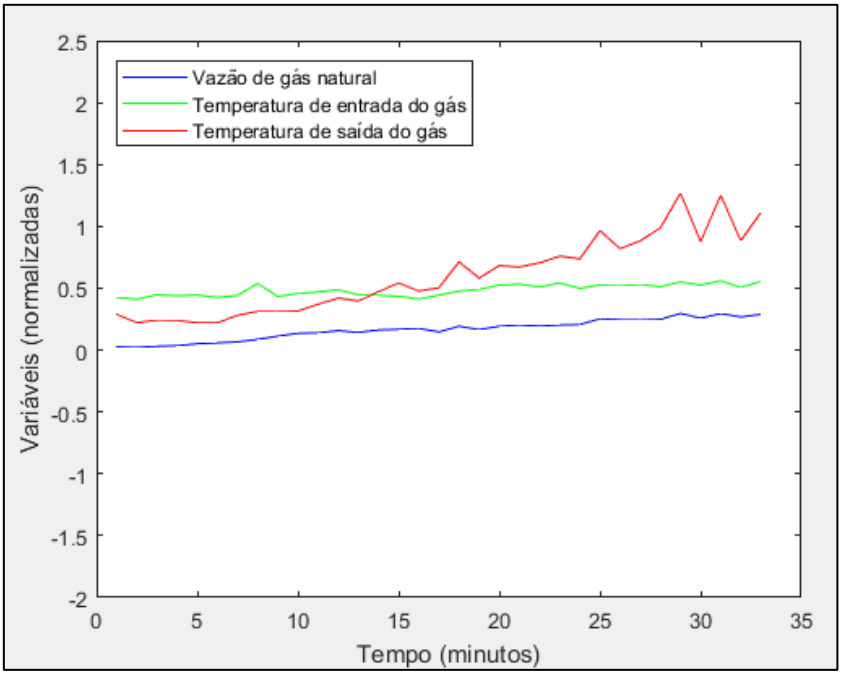


Figura 22 – Padrão normal 1 obtido (FCM-GA).

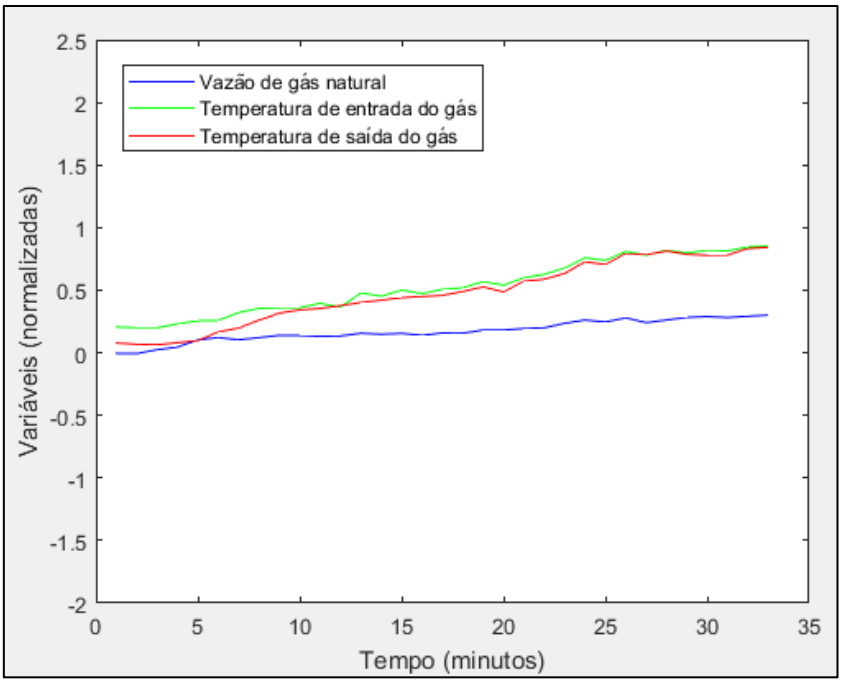


Figura 23 – Padrão normal 2 (FCM-GA).

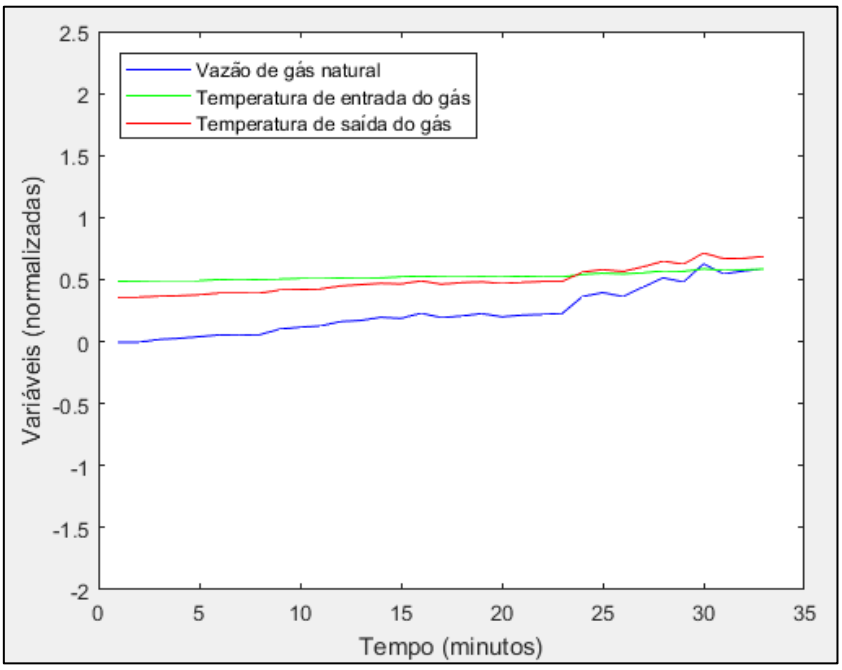


Figura 24 – Padrão de falha (FCM-GA).

A Figura 25 apresenta o gráfico da soma das distâncias entre os centros, distâncias extragrupos (segunda parcela da função de avaliação, eq. 16) que aumenta no decorrer do processo evolutivo até atingir a sua convergência.

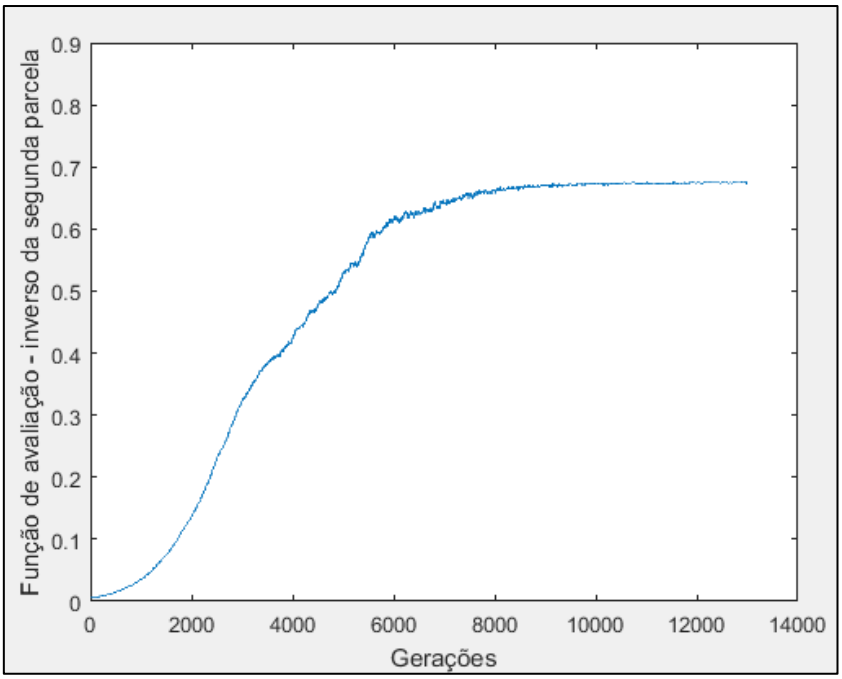


Figura 25 – Soma das distâncias extragrupos (FCM-GA).

Por outro lado, a soma das distâncias intragrupos (primeira parcela da função de avaliação, eq. 16) diminui no decurso das iterações (Figura 26), o que evidencia a elevação da similaridade entre os objetos de um mesmo grupo ao longo das sucessivas gerações.

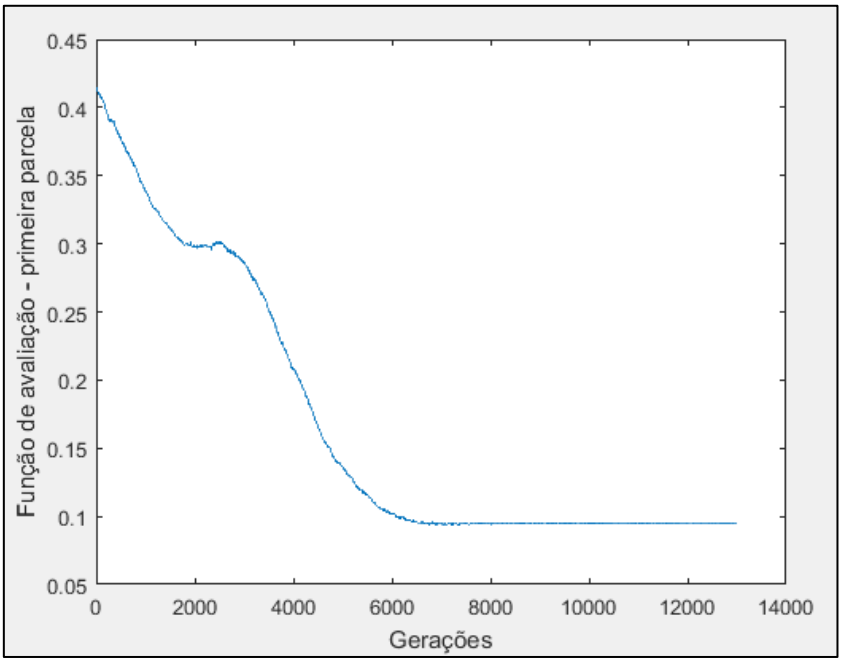


Figura 26 – Soma das distâncias intragrupos (FCM-GA).

No intuito de validar a aplicação do GA ao problema, segundo a metodologia proposta (seção 3.1), o mesmo algoritmo FCM, baseado em um método clássico (não-heurístico) de otimização, foi aplicado ao problema. Foram adotados a mesma função objetivo (eq. 16), a mesma formulação para determinação da matriz de partição (eq. 5), a mesma estimativa inicial e as mesmas amostras (treinamento e teste). Ou seja, tem-se rigorosamente o mesmo problema de otimização, porém com diferentes estratégias de resolução.

O algoritmo genético de agrupamento proposto alcançou melhores resultados quando comparado ao método tradicional de agrupamento (FCM) baseado em um algoritmo clássico de otimização. A tabela 5 apresenta as taxas de erro obtidas com a aplicação do FCM tradicional. Com o FCM-GA, 3 objetos (10%) de partida normal foram classificados de forma errada na amostra de treinamento enquanto no FCM baseado em otimização clássica 5 objetos (16,7%) foram classificados de forma errada.

O percentual de 20% de classificações errada em relação aos objetos de falha na amostra de treinamento equivale a 2 objetos (de um total de 10) de falha que foram classificados como de partida normal. Este mesmo resultado já foi obtido em trabalhos anteriores que propuseram, inclusive, abordagem de métrica híbrida para o FCM (FONTES e BUDMAN, 2017; FONTES e PEREIRA, 2016). Isto sugere a possibilidade de padrões adicionais de falha cujo reconhecimento só seria viabilizado a partir da síntese computacional de outros objetos de falha ou da aquisição de novos dados da UTE. Ou seja, diante da minoria de objetos de falha na amostra disponível, a taxa de erro de 20% parece ser efetivamente o melhor resultado a ser obtido através de uma abordagem de agrupamento não hierárquico FCM.

Tabela 5 – Porcentagem de classificações erradas – FCM com otimização clássica

Amostra	Operação com Falha	Operação Normal
Treinamento	20%	16,7%
Teste	—	3,3%

Um dos três grupos obtidos pelo FCM tradicional não apresentou objetos, mostrando que esta abordagem não foi capaz de reconhecer a existência de um terceiro grupo. Os padrões obtidos pelo FCM tradicional, para cada um dos dois grupos não vazios são apresentados nas Figuras 27 e 28.

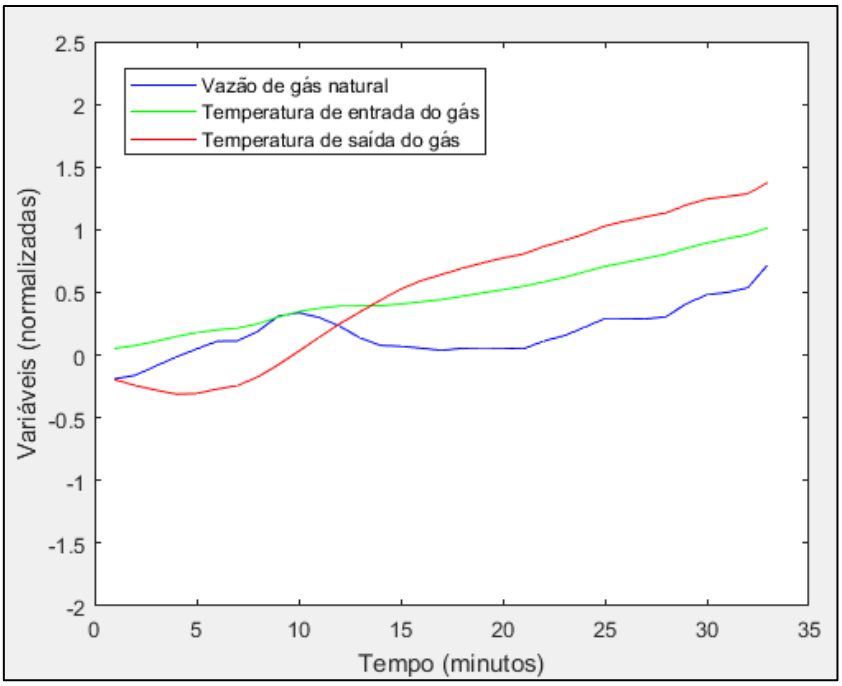


Figura 27 – Padrão normal obtido com o FCM clássico.

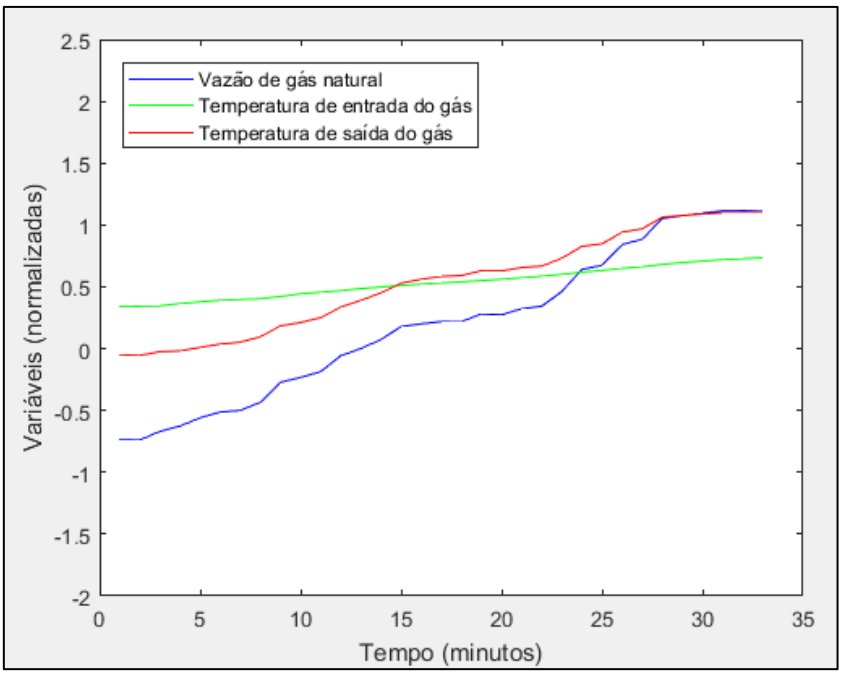


Figura 28 – Padrão de falha obtido com o FCM clássico.

Na Figura 29 é possível visualizar o valor da função objetivo ao longo das iterações.

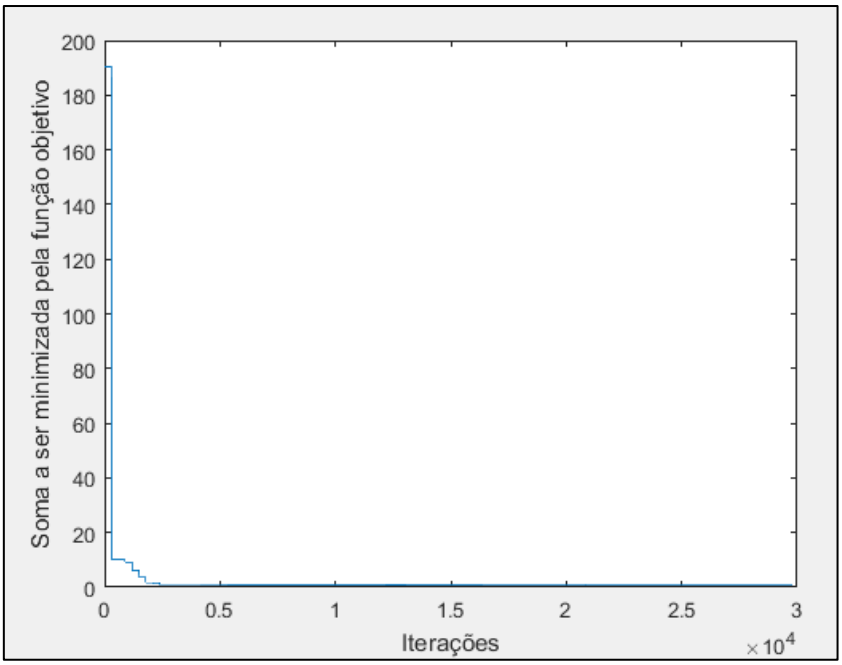


Figura 29 – Função objetivo (FCM com otimização clássica) ao longo das iterações.

Os padrões reconhecidos segundo o FCM-GA (Figuras 22-24) para o grupo normal 1 e para o grupo de falha guardam uma certa semelhança de comportamento dinâmico em relação aos respectivos padrões reconhecidos através do FCM com otimização clássica (Figuras 27 e 28). Por outro lado, as diferenças verificadas atestam a obtenção de soluções (mínimos locais) efetivamente diferentes em ambos os métodos. A Figura 30 apresenta as distâncias ($SPCA_c$) entre os padrões reconhecidos pelo FCM-GA (a) e pelo FCM clássico (b).

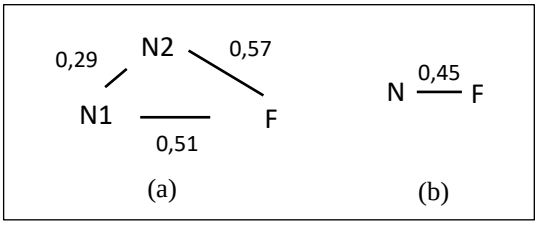


Figura 30 – Distâncias entre os padrões reconhecidos.
(a) FCM-GA. (b) FCM clássico.

Verifica-se (Figura 30(a)) que o padrão N1 (Normal 1) é ligeiramente mais próximo do padrão de falha (F) que o padrão N2 (Normal 2), sugerindo que este último seria um padrão mais seguro (ou mais desejável) de partida da turbina. Além disso, se verifica também que o FCM-GA faz uma distinção entre o comportamento de operação normal e de falha uma vez que os padrões normais estão mais próximos entre si e

relativamente mais distantes, ambos, ao padrão de falha. A Figura 30 (b) mostra que a distância obtida pelo FCM clássico entre os padrões N (Normal) e F (Falha) é menor do que as distâncias N1—F e N2—F obtidas pelo FCM-GA de onde se pode concluir que o que FCM-GA possui padrões normais mais distantes do padrão de falha, o que pode ser considerado como mais conveniente.

4.2 ESTUDO DE CASO 2: DETECÇÃO DE FALHAS EM UMA UNIDADE VIRTUAL

4.2.1 Descrição

O segundo estudo de caso consistiu na análise de um processo virtual de referência, o *Tennessee Eastman Process* (TEP), totalmente disponível na *web*. Embora seja uma planta virtual, o TEP é amplamente adotado na literatura como processo de referência (*benckmark*) para estudos de controle, otimização, monitoramento e análise de falhas (LAU *et al*., 2013; RATO, REIS, 2013; YIN *et al*., 2012; LI *et al*., 2011; LI, XIAO, 2011; ESLAMLOUEYAN, 2011; RICARDEZ-SANDOVAL *et al*., 2009).

O TEP foi desenvolvido pela companhia *Eastman Chemical*, a partir do trabalho de Downs e Vogel (1993), com o objetivo de fornecer um modelo capaz de representar o comportamento de uma planta industrial e possibilitar o desenvolvimento, estudo e avaliação de tecnologias voltadas para o controle e otimização de processos industriais. Tal modelo é baseado em uma planta real, onde os componentes do processo, parâmetros cinéticos e condições operacionais foram modificados por motivos de sigilo de informação (JURICEK *et al*., 2001).

Originalmente desenvolvido em sub-rotinas Fortran, o simulador do processo proposto consiste em cinco operações unitárias: reator, condensador de produto, separador líquido-vapor, compressor de reciclo e uma *stripper*. São produzidos dois produtos a partir de quatro reagentes, além de um subproduto e da presença de uma substância inerte, totalizando oito componentes, sendo eles, A, C, D, E (reagentes), B (inerte), F (subproduto) e G e H (produtos principais). A Figura 31 apresenta um diagrama de todo o processo.

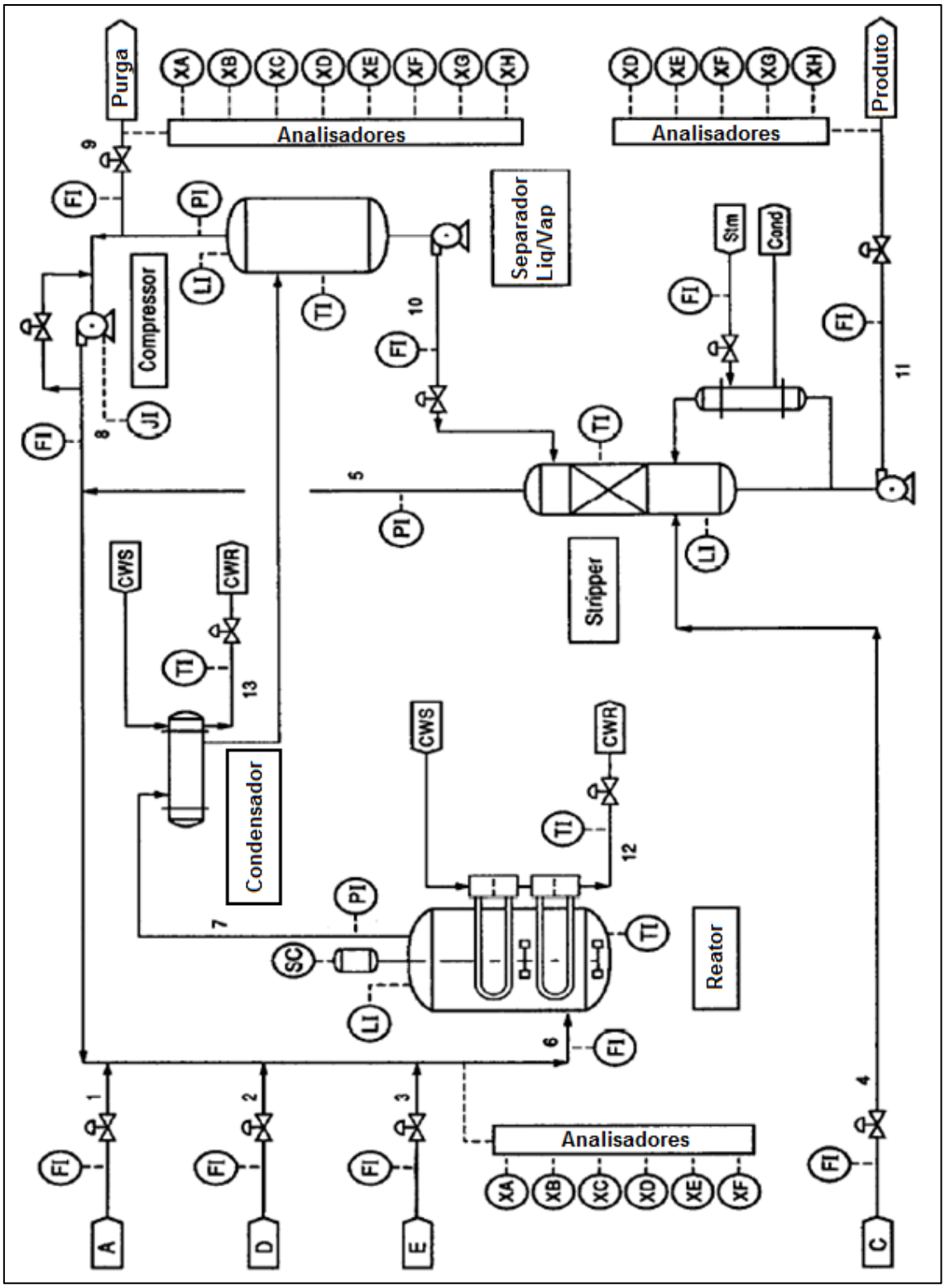


Figura 31 – TEP (DOWNS, VOGEL, 1993).

Inicialmente os reagentes gasosos são alimentados para o reator, onde formam os produtos líquidos. O reator possui um sistema interno de resfriamento para evitar altas temperaturas decorrentes das reações. O produto passa por um condensador e

posteriormente é enviado a um separador líquido-vapor onde os componentes não condensados retornam ao reator através de uma bomba centrífuga de reciclo. Os produtos condensados seguem para uma coluna *stripper*, para remover os reagentes não convertidos do produto final. Então são gerados os produtos G e H na base da *stripper*. Para evitar acúmulo de subproduto e do componente inerte, existe uma corrente de purga (corrente 9).

O processo contém 41 variáveis medidas e mais 12 variáveis manipuláveis. Dentre as variáveis medidas do processo se incluem dados de sensores de temperatura, pressão, nível, vazão das correntes e concentração dos componentes no produto. O simulador do TEP dispõe de 20 possíveis eventos de falha que são essencialmente distúrbios sobre o processo. Estes distúrbios oferecem um potencial conjunto para análise e avaliação de abordagens voltadas à detecção e diagnóstico de falhas. A simulação do TEP viabiliza a obtenção de objetos relacionados a ocorrência de algum tipo de falha (ou mesmo uma combinação entre elas) e dessa forma representa uma alternativa de unidade virtual perfeitamente adequada para a análise e avaliação de estratégias de agrupamento e reconhecimento de padrões (DOWNS, VOGEL, 1993).

A versão original do TEP em malha aberta (Figura 31) é instável. Dentre as alternativas de controle para o TEP propostas na literatura, este trabalho considerou a estratégia apresentada por Ricker (1996) que propõe um controle descentralizado do processo. Este estudo de caso tem sido amplamente utilizado em uma grande variedade de aplicações, certamente por utilizar técnicas de controle multivariável de processos, além de contemplar um modelo em malha fechada já implementado em Matlab/Simulink (Figura 32) (BARRAGAN *et al*., 2016; JURICEK *et al*., 2001).

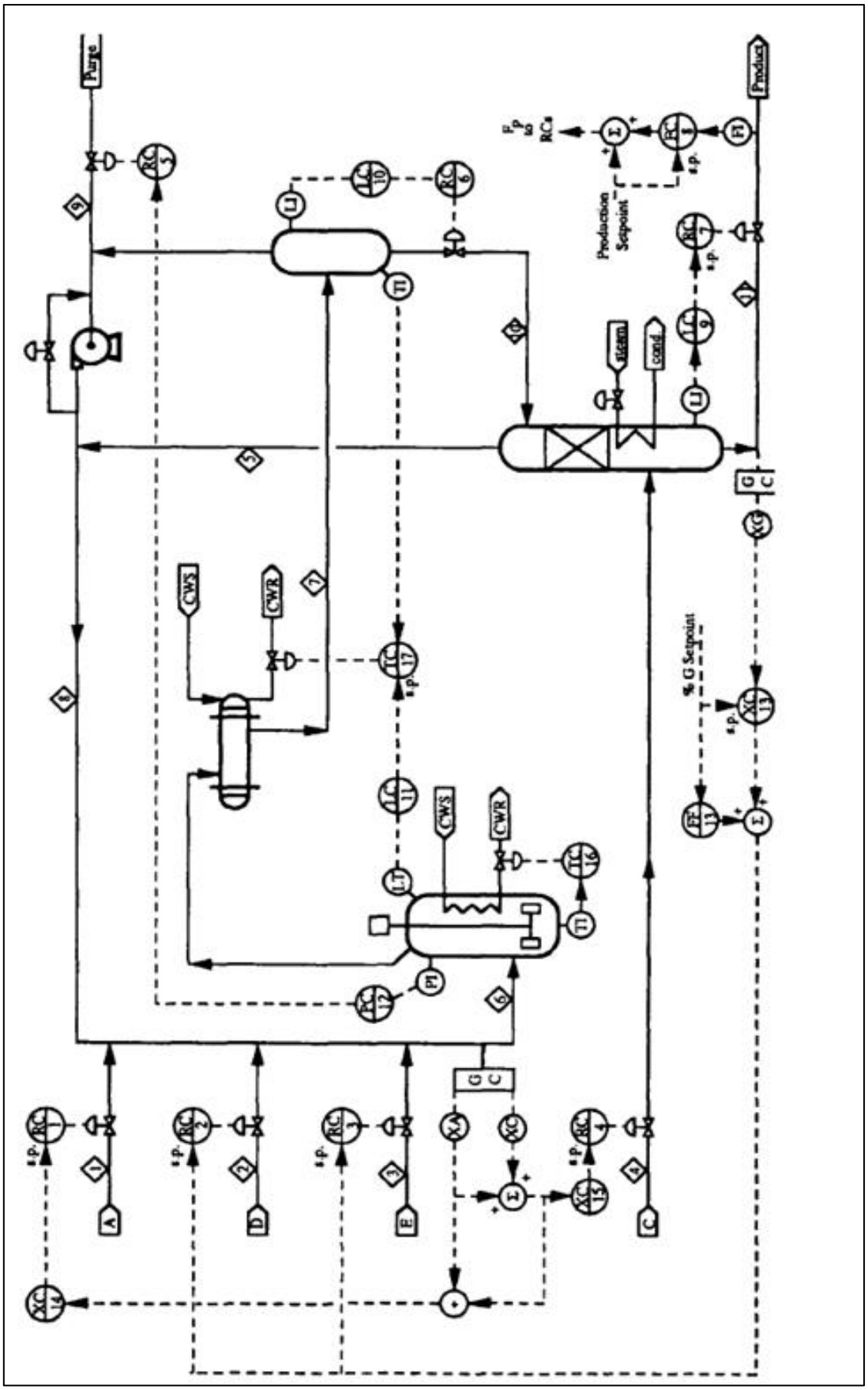


Figura 32 – Sistema de controle aplicado ao TEP (RICKER, 1996).

4.2.2 Coleta de Dados e Resultados

Dentre os vinte tipos de falha definidos no TEP, três são consideradas de difícil detecção ou de observabilidade e resultam elevadas taxas de erros de classificação. Por sua vez, duas destas três falhas (falhas 3 e 9) estão relacionadas a distúrbios na mesma variável de processo (temperatura da corrente de alimentação do reagente D no reator). Assim, a fim de avaliar o desempenho da abordagem proposta na detecção de falhas associadas ao TEP, foram consideradas neste estudo de caso uma falha típica (falha 2) com o mesmo nível de complexidade da maioria das falhas e uma falha adicional com difícil capacidade de detecção (falha 3).

A falha 2 compreende uma perturbação do tipo degrau na composição de inerte (B) mantendo-se a razão entre os reagentes A/C constante (fluxo 4, que é uma mistura de reagentes A e C). Esta falha/distúrbio resulta em mudanças na quantidade de inertes no sistema, na pressão do reator, na pressão parcial dos reagentes dentro do reator e na taxa de produção (produtos G e H). A falha 3 está associada a uma perturbação do tipo degrau na temperatura de alimentação do reagente D (corrente 2). Esta perturbação provoca um efeito quase desprezível (ou imperceptível), em malha fechada, e apenas duas variáveis de processo apresentam uma pequena alteração, quais sejam, temperatura do reator e vazão de líquido de resfriamento na camisa. A dificuldade e os desafios verificados na detecção deste tipo de falha já foram reportados em outros trabalhos da literatura (SHAMS *et al*., 2011; FONTES e BUDMAN, 2018).

A extração dos dados compreendeu inicialmente a simulação do TEP para um tempo total de 2500 h com um período de amostragem de 30 s. A taxa de produção foi randomicamente alterada no intervalo [19,5;26,3] m^3/h. Foi considerada uma falha (2 ou 3) de cada vez e ocorrendo em diferentes estados estacionários (definido pela taxa de produção). Séries temporais multivariadas (objetos) relacionadas a operação normal e de falha foram extraídas considerando-se uma janela de tempo de acordo com a dinâmica de cada falha. Desta forma, para o estudo da falha 2 as séries multivariadas (associadas à falha 2 à operação normal) possuem todas uma janela de tempo de 500 min, enquanto que para a falha 3 as séries compreendem uma janela de tempo de apenas 20 min. Duas amostras de mesmo tamanho (50 objetos de falha e 50 objetos de operação normal) foram consideradas para a análise de cada tipo de falha (2 e 3).

Assim como no primeiro estudo de caso, os grupos e centros foram obtidos usando a amostra de treinamento e, em seguida, os dados de teste foram utilizados para

validar os resultados de classificação. Cada série temporal foi normalizada considerando os valores máximos e mínimos de cada uma das variáveis presentes em toda a amostra (100 objetos para a falha 2 e 100 objetos para a falha 3). Para cada uma das falhas (2 e 3) foram selecionados randomicamente da amostra global 30 objetos normais e 30 de falha para constituir a amostra de treinamento. Os 40 objetos restantes (20 de falha e 20 normais) passaram a compor a amostra de teste. Como este segundo estudo de caso contempla dados sintéticos (obtidos através da simulação de uma unidade virtual), foi possível gerar uma amostra de treinamento balanceada com a mesma quantidade de objetos normais e de falha, ao contrário do primeiro estudo de caso.

Assim como no estudo de caso anterior, adotou-se a mesma estratégia de limitar a taxa de alteração de cada gene na mutação em ±1% em relação ao valor atual para evitar a obtenção de centros com oscilações excessivas. Além disso, da mesma forma que no primeiro estudo de caso, o percentual de classificações incorretas para eventos normais e para eventos com falha obtidos pelo algoritmo genético foi comparado com resultados obtidos pelo FCM usando um método de otimização clássica.

4.2.2.1 Falha 2

Na falha 2 cada objeto é uma série multivariada com três variáveis. As três variáveis de processo consideradas para o estudo são: vazão do reagente D (corrente 2), vazão do reagente E (corrente 3) e vazão do reagente C (corrente 4). A Figura 33 apresenta todas as séries da amostra de treinamento, com a ilustração dos perfis das 3 variáveis analisadas.

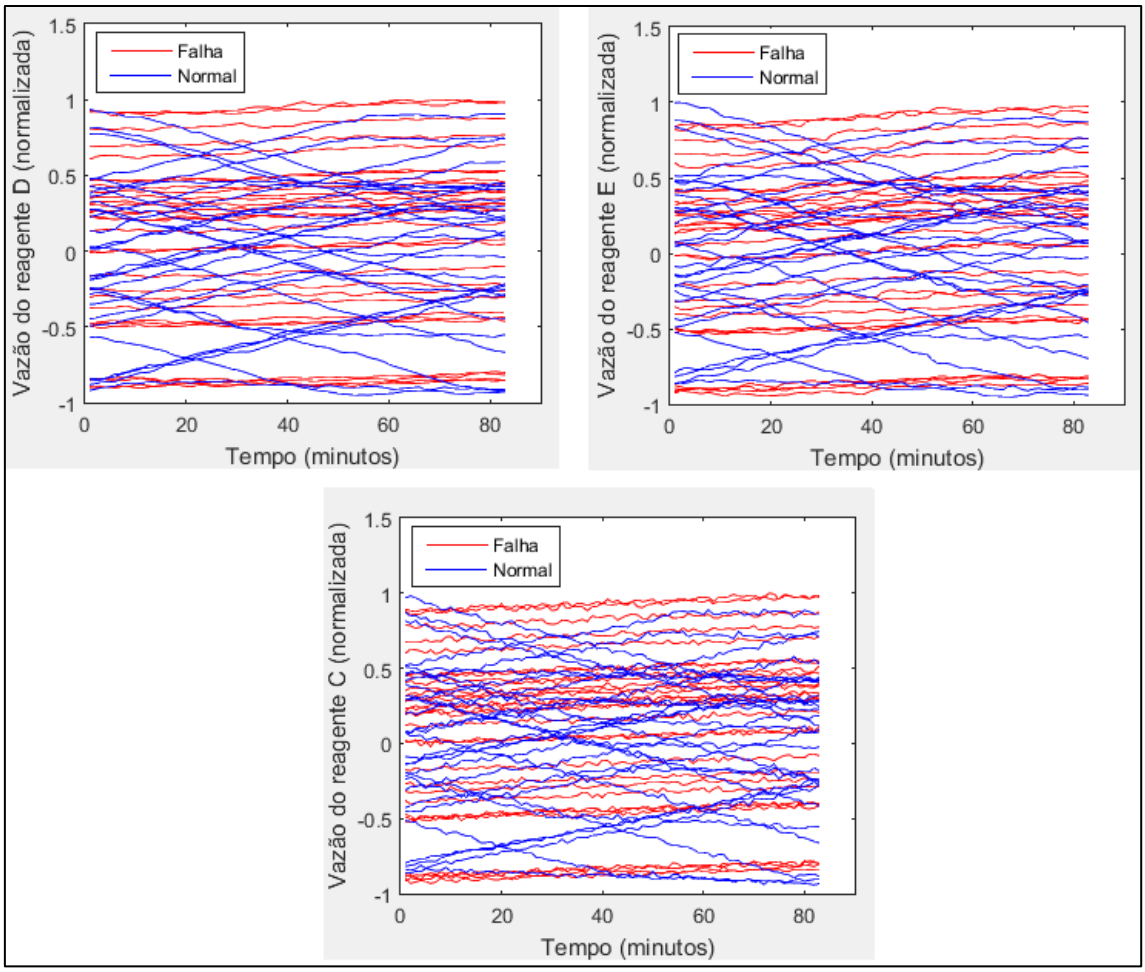


Figura 33 – Séries da amostra de treinamento – TEP falha 2.

Da mesma forma que no estudo de caso anterior, três grupos foram considerados (*c=3*) baseando-se em testes já realizados com outras quantidades de grupos (FONTES e BUDMAN, 2018). A Figura 34 apresenta o resultado final de classificação dos 60 objetos da amostra de treinamento.

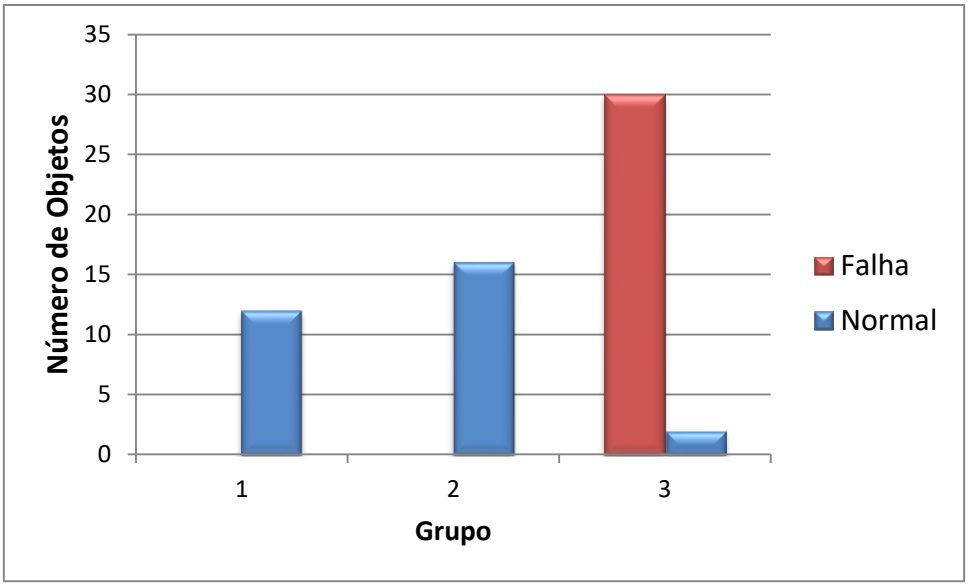


Figura 34 – Objetos distribuídos nos três grupos – TEP falha 2.

Verifica-se que nos grupos 1 e 2 todos os objetos são normais enquanto que o grupo 3 concentrou todos os objetos de falha. Isso sugere que o grupo 3 seja categorizado como tipicamente um grupo de falha. Por sua vez, os grupos 1 e 2 podem ser caracterizados como grupos de objetos normais, o que evidencia também existência de mais de um padrão associado à operação normal, além de um excelente resultado de agrupamento fornecido pelo FCM-GA.

As taxas finais de erro das classificações do FCM-GA, tanto na amostra de treinamento quanto na amostra de teste, para as operações normais e com falha encontram-se na Tabela 6.

Tabela 6 – Porcentagem de classificações erradas (TEP falha 2) (FCM-GA)

Amostra	Operação com Falha	Operação Normal
Treinamento	0%	6,66%
Teste	0%	15%

Os padrões obtidos pelo GA para cada um dos três grupos (dois normais e um com falha) estão apresentados nas Figuras 35-37.

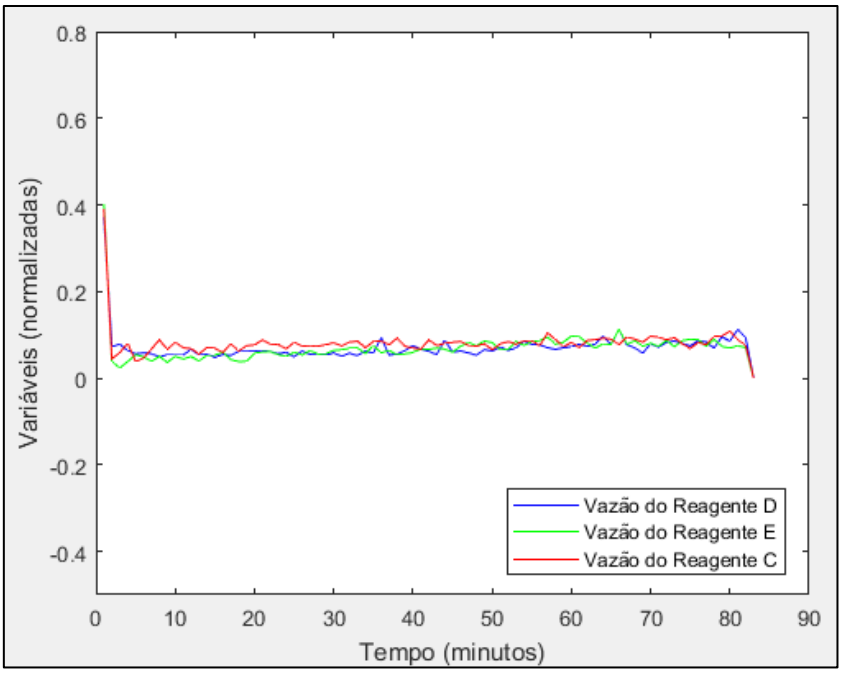


Figura 35 – Padrão normal 1 obtido (FCM-GA) – TEP falha 2.

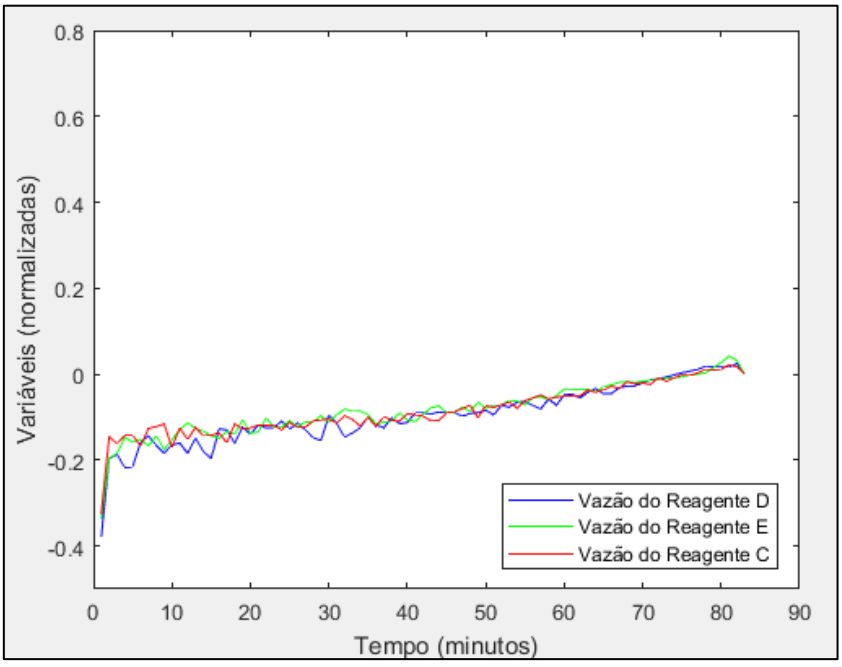


Figura 36 – Padrão normal 2 obtido (FCM-GA) – TEP falha 2.

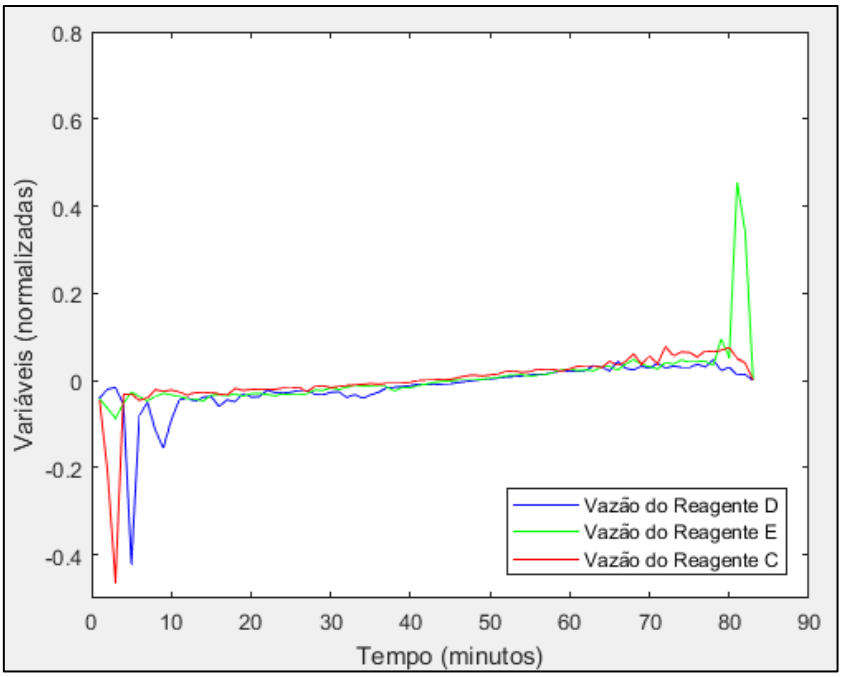


Figura 37 – Padrão de falha obtido (FCM-GA) – TEP falha 2.

O algoritmo genético de agrupamento proposto alcançou melhores resultados quando comparado ao método tradicional de agrupamento, FCM baseado em um algoritmo clássico de otimização. A tabela 7 apresenta as taxas de erro das classificações do FCM tradicional para as mesmas amostras de treinamento e teste submetidas ao GA.

Tabela 7 – Porcentagem de classificações erradas (TEP falha 2) – otimização clássica

Amostra	Operação com Falha	Operação Normal
Treinamento	0%	23,33%
Teste	0%	20%

Os padrões obtidos pelo FCM com otimização clássica podem ser visualizados nas Figuras 38-40.

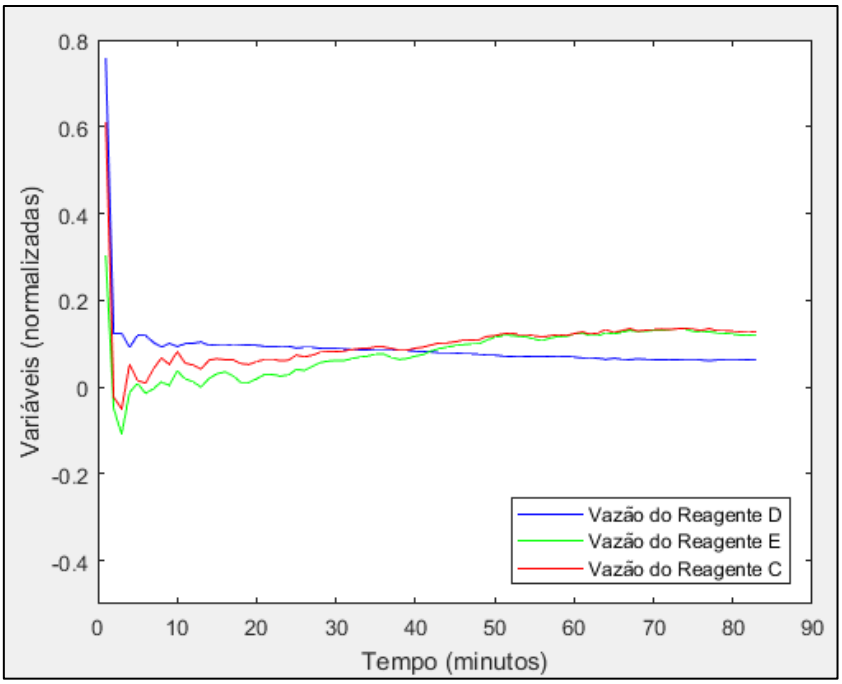


Figura 38 – Padrão normal 1 obtido com o FCM clássico – TEP falha 2.

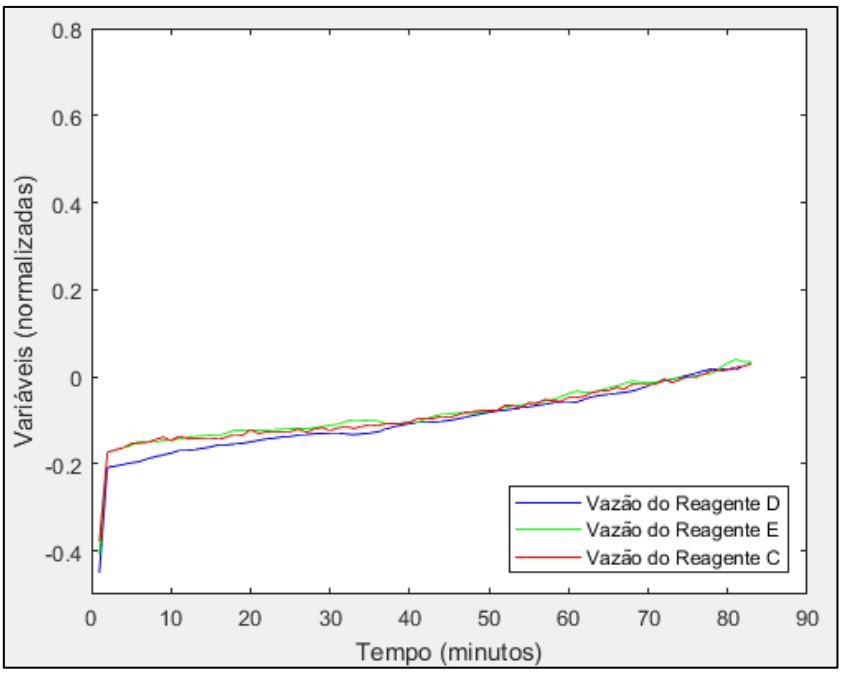


Figura 39 – Padrão normal 2 obtido com o FCM clássico – TEP falha 2.

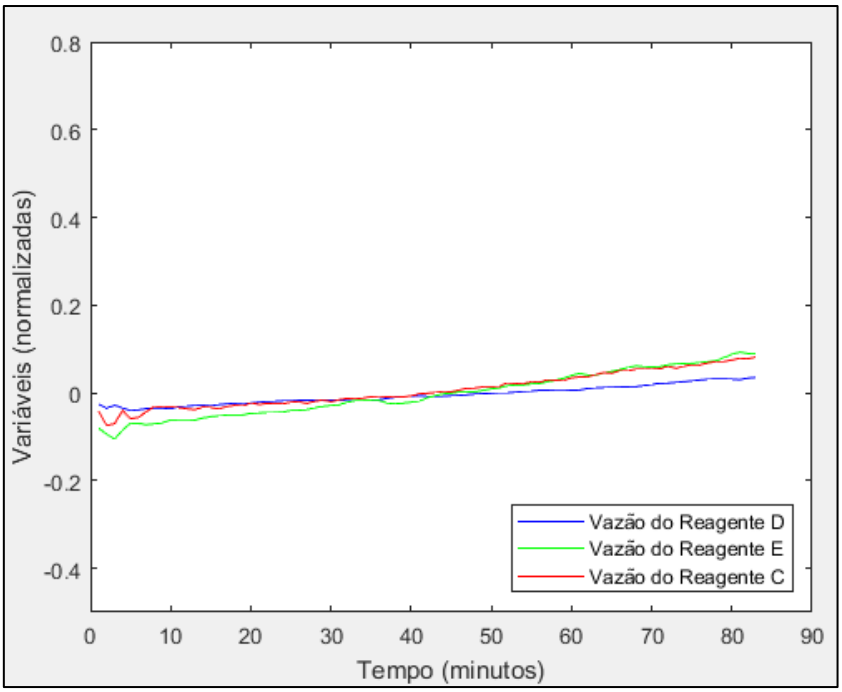


Figura 40 – Padrão de falha obtido com o FCM clássico – TEP falha 2.

Confrontando as tabelas 6 e 7 é possível perceber que o algoritmo genético obteve (na operação normal) uma taxa de erro bastante inferior ao método de FCM com otimização clássica, o que evidencia de fato a obtenção de um melhor mínimo local para o primeiro, mesmo considerando-se que em ambas as abordagens a taxa de erro de classificação não foi adotada explicitamente como critério de minimização (função objetivo) ou de avaliação de aptidão/qualidade da solução. Além disso, reitera-se o fato de que o algoritmo de agrupamento compreende essencialmente um aprendizado não supervisionado no qual o rótulo de cada objeto (falha ou normal) não é informado no processo de otimização.

Os resultados obtidos neste segundo estudo de caso (e também no primeiro) ratificam a eficácia da metodologia de aplicação de algoritmos genéticos para o agrupamento de séries temporais multivariadas usando o método FCM.

Tal como no estudo de caso anterior, os padrões reconhecidos por ambas as abordagens (FCM-GA e FCM com otimização clássica) apresentam comportamentos dinâmicos semelhantes. Por outro lado, os picos (máximos e mínimos) verificados em ambos os casos sugerem uma questão adicional que se refere à factibilidade ou viabilidade de realização destes padrões no processo analisado. Este trabalho não abordou esta questão e não foram adicionados ao problema de otimização restrições

(suaves ou severas) com o objetivo de reconciliar ao aproximar os padrões reconhecidos à realidade do processo. O objetivo prioritário dos padrões reconhecidos é a viabilidade de obtenção de partição satisfatória dos objetos entre grupos de falha e de operação normal.

A Figura 41 apresenta as distâncias ($SPCA_c$) entre os padrões reconhecidos para a falha 2 pelo FCM-GA (a) e pelo FCM clássico (b).

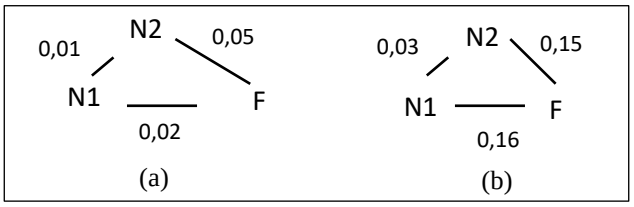


Figura 41 – Distâncias entre os padrões reconhecidos – TEP falha 2.
(a) FCM-GA. (b) FCM clássico.

A Figura 41 mostra que em ambas as abordagens (GA e otimização clássica) foi possível obter padrões de operação mais próximos entre si e ambos mais distantes ao padrão de falha. Apesar de um resultado de classificação menos satisfatório (Tabela 7), o FCM com otimização clássica obteve padrões de operação normal mais distantes do comportamento de falha.

Adicionalmente aos testes realizados, optou-se por alterar a estimativa inicial (centros iniciais) em ambas as abordagens, ou seja, causar uma perturbação na entrada de modo a verificar o comportamento dos algoritmos. O FCM-GA convergiu para uma taxa de classificação satisfatória. Já o FCM baseado em otimização clássica colocou quase todas as séries em um mesmo grupo, resultando numa elevada taxa de erro, o que demonstra a sua falta de robustez (tabelas 8 e 9).

Tabela 8 – Porcentagem de classificações erradas (TEP falha 2 – alteração na estimativa inicial) – FCM-GA

Amostra	Operação com Falha	Operação Normal
Treinamento	3,33%	3,33%
Teste	0%	5%

Tabela 9 – Porcentagem de classificações erradas (TEP falha 2 – alteração na estimativa inicial) – otimização clássica

Amostra	Operação com Falha	Operação Normal
Treinamento	0%	93,33%
Teste	0%	90%

4.2.2.2 Falha 3

No caso da falha 3 cada objeto é uma série multivariada com duas variáveis, quais sejam, vazão da água de resfriamento na camisa do reator e temperatura do reator (as únicas variáveis medidas que apresentam alguma alteração não desprezível devido à ocorrência da falha). As Figuras 42 e 43 apresentam algumas séries temporais relacionadas a alguns objetos normais e de falha considerando as duas variáveis de processo selecionadas. Mesmo após a obtenção de séries com uma janela de tempo de 20 min (período de amostragem de 30 s), verificou-se que a redução da janela para 7,8 min ofereceu melhores resultados de classificação. A janela de 7,8 min viabilizou uma melhora na informação contida nos perfis no sentido de reconhecer melhor as dissimilaridades entre os comportamentos de falha e de operação normal, segundo as métricas envolvidas (SPCA e AED) (FONTES e BUDMAN, 2018).

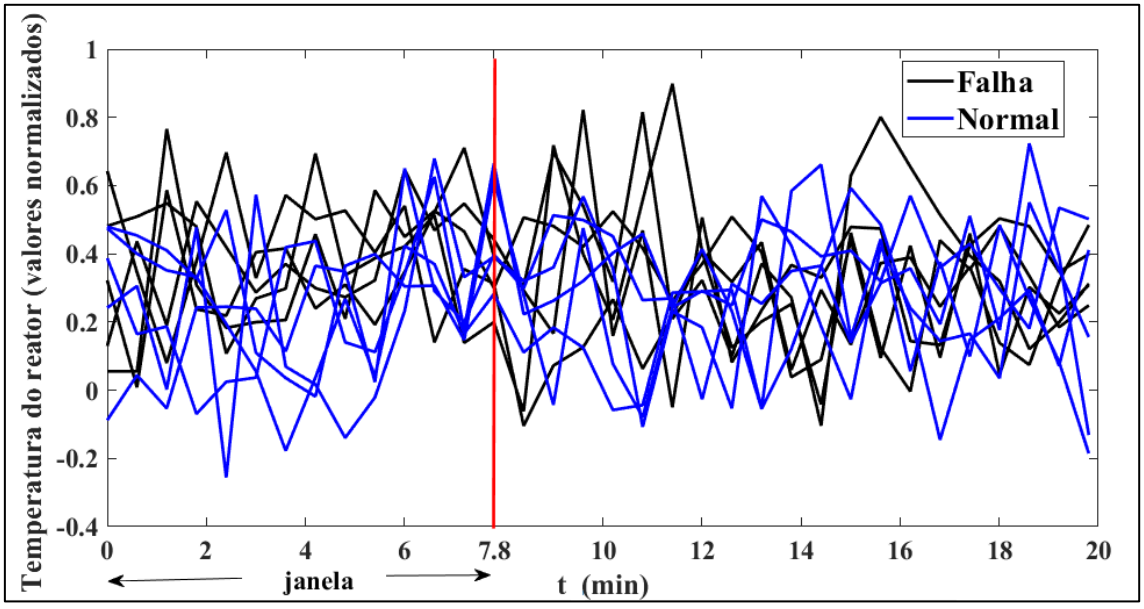


Figura 42 – Algumas séries da amostra (variável: temperatura).

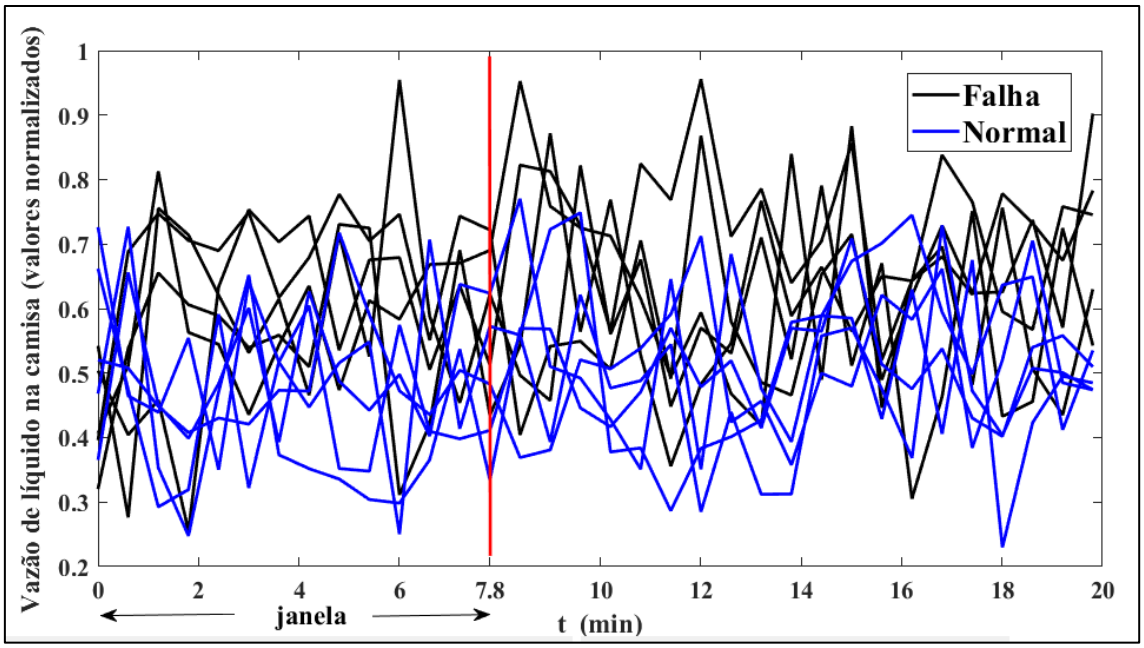


Figura 43 – Algumas séries da amostra (variável: vazão de líquido).

É possível perceber a partir das Figuras 42 e 43 que o comportamento das variáveis de processo nas séries temporais verificadas apresenta uma elevada similaridade, inclusive do ponto de vista de direção de variabilidade o que evidencia a dificuldade de reconhecimento de um padrão de falha para este problema.

Dois grupos foram considerados. O resultado obtido foi a segregação das 60 séries multivariadas da amostra de treinamento, originando um grupo de curvas normais e outro de curvas com falha com seus respectivos padrões (Figuras 44 e 45).

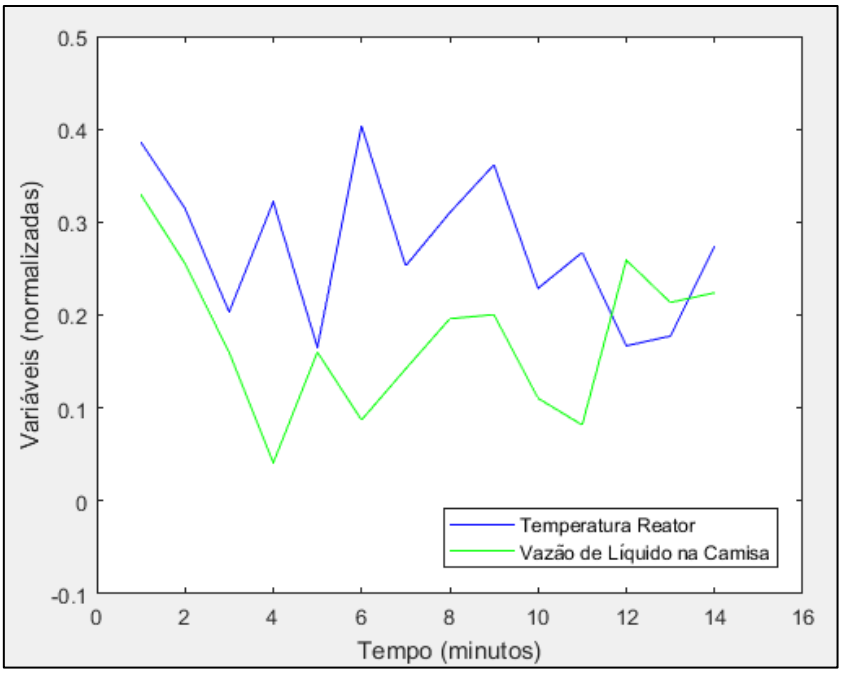


Figura 44 – Padrão normal obtido com o GA – TEP falha 3.

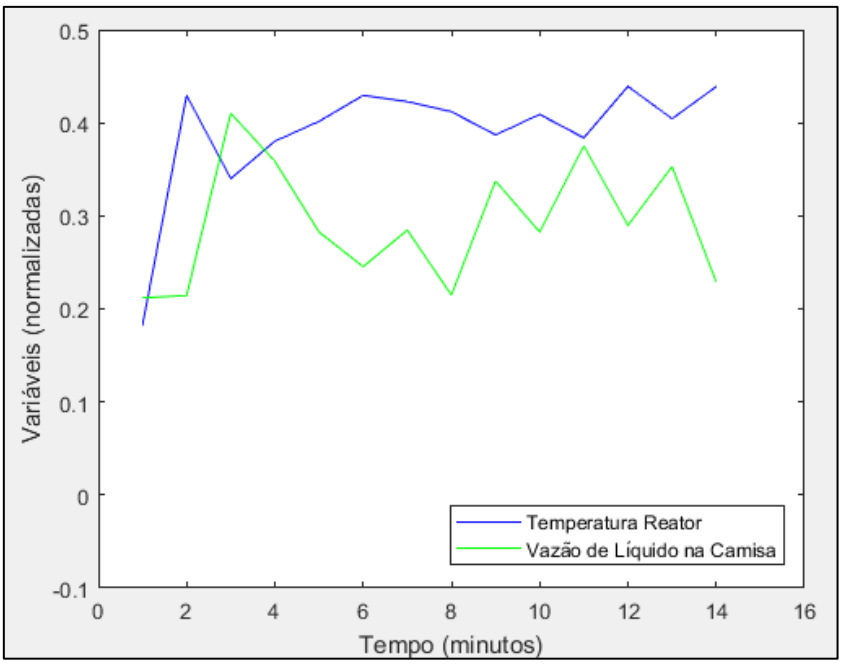


Figura 45 – Padrão de falha obtido com o GA – TEP falha 3.

As taxas de erro de classificação obtidas pelo FCM-GA e pelo FCM com otimização clássica foram exatamente as mesmas, tanto na amostra de treinamento quanto na amostra de teste, para as operações normais e com falha (Tabela 10).

Tabela 10 – Porcentagem de classificações erradas (TEP falha 3) – GA/FCM

Amostra	Operação com Falha	Operação Normal
Treinamento	26,66%	26,66%
Teste	15%	15%

Os resultados obtidos reforçam a limitação de se obter melhores resultados com a falha 3. Torna-se evidente a dificuldade de detecção do tipo de falha considerado o que, por sua vez, estaria muito mais associado à pouca informação nos dados do que propriamente ao nível de complexidade do espaço de busca. Além disso, os resultados encontrados estão dentro do limite de aceitabilidade verificado em outros trabalhos da literatura que tratam deste mesmo tipo de falha (FONTES e BUDMAN, 2018; SHAMS *et al*., 2011). O fato do FCM-GA ter obtido o mesmo resultado do FCM baseado em otimização clássica apenas reforça a robustez do primeiro em, pelo menos, não ter alcançado um mínimo local de qualidade inferior.

Os padrões obtidos pelo FCM tradicional estão apresentados nas Figuras 46 e 47 e são bastante similares aos padrões obtidos pelo FCM-GA (já apresentados nas Figuras 44 e 45), o que está consistente com o fato das classificações terem sido exatamente as mesmas, em ambos os algoritmos de otimização.

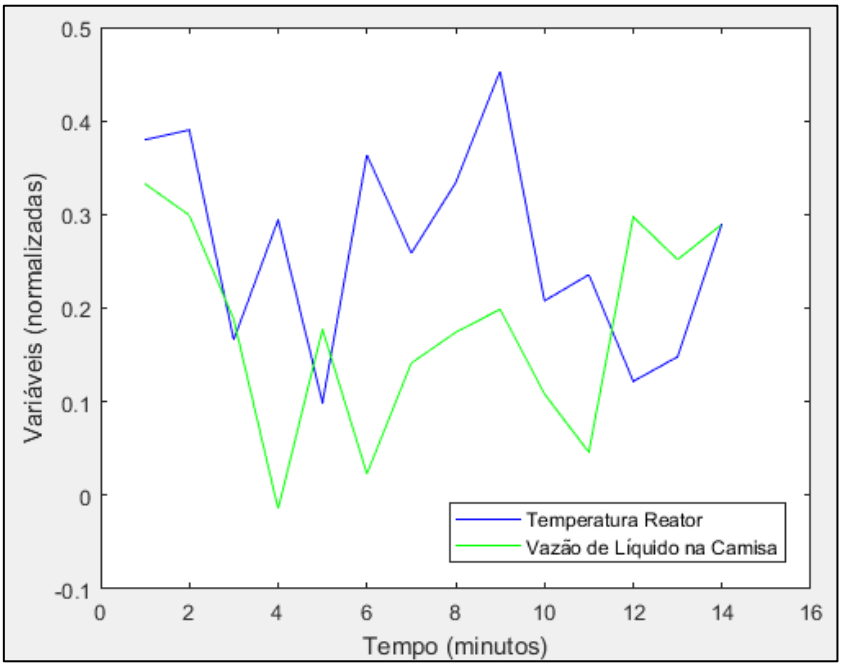


Figura 46 – Padrão normal obtido com o FCM otimização clássica – TEP falha 3.

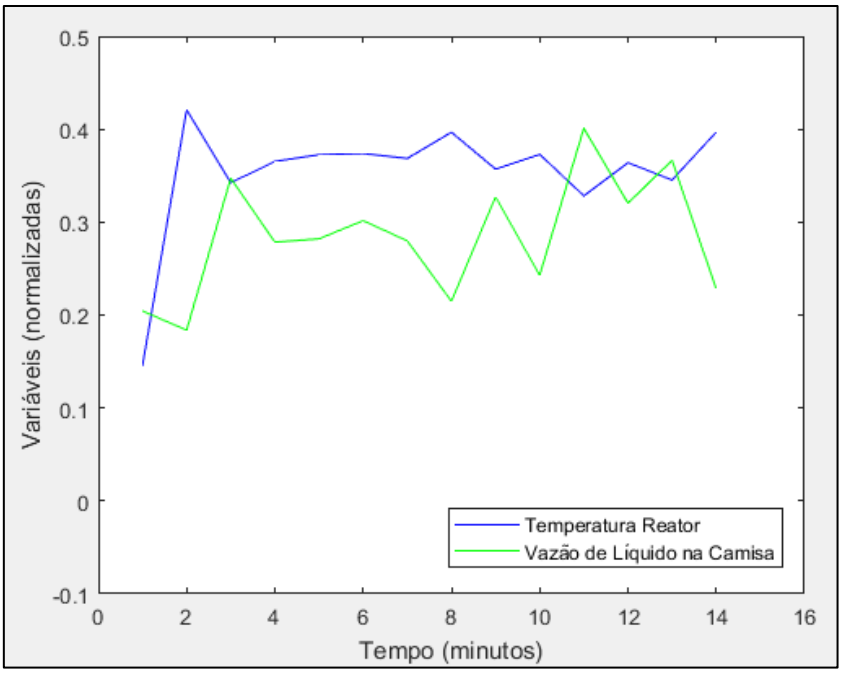


Figura 47 – Padrão de falha obtido com o FCM otimização clássica – TEP falha 3.

CAPÍTULO 5

CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

5.1 CONCLUSÕES

Uma metodologia para a aplicação de algoritmos genéticos (GA) no agrupamento e reconhecimento de padrões em séries temporais multivariadas é proposta neste trabalho. Pela primeira vez é proposta uma abordagem que viabiliza e sugere uma alternativa de adequação do GA no contexto do FCM tendo-se objetos que são representados por matrizes (séries multivariadas). O procedimento compreende, entre outros, a definição da codificação do cromossomo (estrutura em duas dimensões) e a customização dos operadores intrínsecos ao GA (seleção, *crossover* e mutação) para o problema analisado, além da métrica de aptidão (função objetivo) de cada indivíduo em uma dada população.

Dois estudos de caso foram analisados. No primeiro, utilizou-se uma base de dados extraída de uma unidade de operação real e os resultados obtidos demonstraram que o agrupamento FCM-GA foi capaz de classificar e reconhecer padrões satisfatoriamente, considerando as limitações da amostra inerentes às informações e dados disponíveis. O segundo estudo de caso compreendeu um processo industrial virtual de referência (*benchmark*) (*Tennessee Eastman Process*) que é amplamente utilizado para a avaliação de abordagens de controle e detecção de falhas.

O agrupamento baseou-se na utilização de uma métrica de similaridade apropriada aos objetos (SPCA, ou PCA *Similarity fator*), de forma conjugada ou não a outras métricas também propostas na literatura. No intuito de validar o método proposto, o algoritmo FCM tradicional foi adaptado para tratar séries temporais multivariadas, visto que este originalmente admite apenas séries temporais univariadas.

Nos dois estudos de caso, verificou-se um aumento no percentual de acerto de classificação, quando se compara o FCM-GA ao método FCM baseado em otimização clássica o que comprova a maior eficácia do primeiro na tentativa de explorar substancialmente a região de busca e obter um melhor resultado de mínimo local. Ou seja, o que se verifica como potencialidade dos métodos heurísticos, em relação à otimização clássica, em problemas diversos (inclusive séries univariadas), também se constata para o agrupamento e reconhecimento de padrões em séries multivariadas. O FCM-GA mostrou-se mais estável (robusto) em relação a escolha do chute (estimativa) inicial, ao contrário do FCM baseado em otimização clássica.

Abordando um tema de pesquisa bastante recente, onde ainda há um número reduzido de pesquisas científicas abordando o assunto, os resultados obtidos trazem importantes contribuições quanto à utilização dos algoritmos genéticos aplicados ao reconhecimento de padrões, especialmente em séries temporais multivariadas.

Os padrões reconhecidos, quando factíveis ou viáveis de realização no processo, constituem-se em referência importantes para tomada de decisão, sobretudo em problemas de detecção de falhas, permitindo o monitoramento em tempo real do processo em relação à sua distância ao comportamento de normalidade ou anormalidade.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

- Incorporar ao FCM baseado em GA restrições adicionais que viabilizem a reconciliação de padrões reconhecidos ou mesmo a maior proximidade em relação à realidade do processo;
- Proposição de FCM-GA com uma abordagem de otimização multi-objetivo;
- Implementar o FCM com outros métodos de otimização heurística.

5.3 PUBLICAÇÃO

RIBEIRO, K. P; FONTES, C. H. O. A Genetic Algorithm Based Clustering Applied to Multivariate Time Series. In: 24TH ABCM INTERNATIONAL CONGRESS OF MECHANICAL ENGINEERING, 2017. Curitiba, PR, Brazil. **Anais**... ABCM, 2017.

REFERÊNCIAS

ABONYI, Janos *et al*. Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series. **Fuzzy Sets and Systems**, v. 149, n. 1, p. 39-56, 2005.

ABOUSLEIMAN, Rami. Roulette Wheel Selection. **Mathworks**, United States, 2015. Disponível em: <https://www.mathworks.com/matlabcentral/fileexchange/45735-roulette-wheel-selection>. Acesso em: 20 out. 2018.

AGGARWAL, Charu C.; REDDY, Chandan K. (Ed.). **Data clustering: algorithms and applications**. CRC press, 2013.

AHONEN, Timo; HADID, Abdenour; PIETIKAINEN, Matti. Face description with local binary patterns: Application to face recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 28, n. 12, p. 2037-2041, 2006.

ALAEI, Hesam Komari; SALAHSHOOR, Karim; ALAEI, Hamed Komari. A new integrated on-line fuzzy clustering and segmentation methodology with adaptive PCA approach for process monitoring and fault detection and diagnosis. **Soft Computing**, v. 17, n. 3, p. 345-362, 2013.

ALBUS, John Edward *et al*. **Syntactic pattern recognition, applications**. Springer Science & Business Media, 2012.

ARAÚJO, Sidnei Alves de. **Casamento de padrões em imagens digitais livre de segmentação e invariante sob transformações de similaridade**. 2009. Tese de Doutorado. Universidade de São Paulo.

AREL, Itamar; ROSE, Derek C.; KARNOWSKI, Thomas P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. **IEEE computational intelligence magazine**, v. 5, n. 4, p. 13-18, 2010.

BANDYOPADHYAY, Sanghamitra; MAULIK, Ujjwal. Genetic clustering for automatic evolution of clusters and application to image classification. **Pattern recognition**, v. 35, n. 6, p. 1197-1208, 2002.

BANKÓ, ZoltáN; ABONYI, JáNos. Correlation based dynamic time warping of multivariate time series. **Expert Systems with Applications**, v. 39, n. 17, p. 12814-12823, 2012.

BARAGONA, Roberto. A simulation study on clustering time series with metaheuristic methods. **Quaderni di Statistica**, v. 3, p. 1-26, 2001.

BARRAGAN, J. F. M. **Contribuições da Transformada Wavelet para o Agrupamento e Reconhecimento de Padrões em Séries Temporais**. Dissertação

(Mestrado em Engenharia Industrial) – PEI, Universidade Federal da Bahia, Salvador, 2016.

BARRAGAN, João Francisco; FONTES, Cristiano Hora; EMBIRUÇU, Marcelo. A wavelet-based clustering of multivariate time series using a multiscale SPCA approach. **Computers & Industrial Engineering**, v. 95, p. 144-155, 2016.

BELLMAN, Richard; KALABA, Robert; ZADEH, L. Abstraction and pattern classification. **Journal of Mathematical Analysis and Applications**, v. 13, n. 1, p. 1-7, 1966.

BERSON, Alex; SMITH, Stephen J. **Building data mining applications for CRM**. McGraw-Hill, Inc., 2002.

BEZDEK, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms Plenum, New York,(1981).

BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

BEZDEK, James C.; HALL, L. O.; CLARKE, L_P. Review of MR image segmentation techniques using pattern recognition. **Medical Physics**, v. 20, n. 4, p. 1033-1048, 1993.

BEZDEK, James C.; HATHAWAY, Richard J. Optimization of fuzzy clustering criteria using genetic algorithms. In: **Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on**. IEEE, 1994. p. 589-594.

BEZDEK, James C. *et al*. **Fuzzy models and algorithms for pattern recognition and image processing**. Springer Science & Business Media, 2006.

BEZDEK, James C. **Pattern recognition with fuzzy objective function algorithms**. Springer Science & Business Media, 2013.

CHEN, Jingfeng et al. Textual analysis and visualization of research trends in data mining for electronic health records. **Health Policy and Technology**, v. 6, n. 4, p. 389-400, 2017.

CHIANG, Leo H.; RUSSELL, Evan L.; BRAATZ, Richard D. **Fault detection and diagnosis in industrial systems**. Springer Science & Business Media, 2000.

CHIOU, Yu-Chiun; LAN, Lawrence W. Genetic clustering algorithms. **European Journal of Operational Research**, v. 135, n. 2, p. 413-427, 2001.

DASH, Sourabh; VENKATASUBRAMANIAN, Venkat. Challenges in the industrial applications of fault diagnostic systems. **Computers & Chemical Engineering**, v. 24, n. 2, p. 785-791, 2000.

DE OLIVEIRA, José Valente; PEDRYCZ, Witold (Ed.). **Advances in Fuzzy Clustering and its Applications**. John Wiley & Sons, 2007.

DELGADO, Miguel; GÓMEZ-SKARMETA, Antonio Fernandez; MARTÍN, Fernando. A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling. **IEEE Transactions on Fuzzy Systems**, v. 5, n. 2, p. 223-233, 1997.

DENG, Xiaogang; TIAN, Xuemin. Multivariate statistical process monitoring using multi-scale kernel principal component analysis. **In: Fault Detection, Supervision and Safety of Technical Processes 2006**. 2007. p. 108-113.

DENG, Xiaogang; TIAN, Xuemin. Nonlinear process fault pattern recognition using statistics kernel PCA similarity factor. **Neurocomputing**, v. 121, p. 298-308, 2013.

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

DUNN, Joseph C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 1973.

D'URSO, Pierpaolo; MAHARAJ, Elizabeth Ann. Wavelets-based clustering of multivariate time series. **Fuzzy Sets and Systems**, v. 193, p. 33-61, 2012.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados: fundamentos e aplicações**. São Paulo: Pearson Addison Wesley, 2005.

ESLAMLOUEYAN, Reza. Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process. **Applied Soft Computing**, v. 11, n. 1, p. 1407-1415, 2011.

FIRSCHEIN, O. Syntactic pattern recognition and applications. **Proceedings of the IEEE**, v. 71, n. 10, p. 1231-1231, 1983.

FONTES, Cristiano Hora; BUDMAN, Hector. A hybrid clustering approach for multivariate time series–A case study applied to failure analysis in a gas turbine. **ISA transactions**, v. 71, p. 513-529, 2017.

FONTES, Cristiano Hora; BUDMAN, Hector Marcelo. Evaluation of a Hybrid Clustering Approach for a Benchmark Industrial System. **Industrial & Engineering Chemistry Research**, 2018.

FONTES, Cristiano Hora; PEREIRA, Otacílio. Pattern recognition in multivariate time series–A case study applied to fault detection in a gas turbine. Engineering **Applications of Artificial Intelligence**, v. 49, p. 10-18, 2016.

FU, Tak-chung. A review on time series data mining. **Engineering Applications of Artificial Intelligence**, v. 24, n. 1, p. 164-181, 2011.

GARAI, Gautam; CHAUDHURI, B. B. A novel genetic algorithm for automatic clustering. **Pattern Recognition Letters**, v. 25, n. 2, p. 173-187, 2004.

GOLDBERG, David E. Genetic algorithms in search, optimization, and machine learning, 1989. **Reading: Addison-Wesley**, 1989.

HALL, Lawrence O.; OZYURT, Ibrahim Burak; BEZDEK, James C. Clustering with a genetically optimized approach. **IEEE Transactions on Evolutionary computation**, v. 3, n. 2, p. 103-112, 1999.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

HAYKIN, Simon. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

HOLLAND, John H. Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence. **Ann Arbor, MI: University of Michigan Press**, 1975.

HOUCK, Christopher R.; JOINES, Jeff; KAY, Michael G. A genetic algorithm for function optimization: a Matlab implementation. **North Carolina State University**, v. 95, n. 09, 1995.

HRUSCHKA, Eduardo Raul *et al*. A survey of evolutionary algorithms for clustering. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 39, n. 2, p. 133-155, 2009.

HUANG, Yunbing; MCAVOY, Thomas J.; GERTLER, Janos. Fault isolation in nonlinear systems with structured partial principal component analysis and clustering analysis. **The Canadian Journal of Chemical Engineering**, v. 78, n. 3, p. 569-577, 2000.

ISERMANN, Rolf; BALLÉ, Peter. Trends in the application of model-based fault detection and diagnosis of technical processes. **Control Engineering Practice**, v. 5, n. 5, p. 709-719, 1997.

ISERMANN, Rolf. Model-based fault-detection and diagnosis–status and applications. **Annual Reviews in Control**, v. 29, n. 1, p. 71-85, 2005.

IZAKIAN, Hesam; ABRAHAM, Ajith; SNÁŠEL, Václav. Fuzzy clustering using hybrid fuzzy c-means and fuzzy particle swarm optimization. In: **Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on**. IEEE, 2009. p. 1690-1694.

IZAKIAN, Hesam; PEDRYCZ, Witold; JAMAL, Iqbal. Fuzzy clustering of time series data using dynamic time warping distance. **Engineering Applications of Artificial Intelligence**, v. 39, p. 235-244, 2015.

JAIN, Anil K.; DUBES, Richard C. **Algorithms for clustering data**. Prentice-Hall, Inc., 1988.

JAIN, Anil K.; MURTY, M. Narasimha; FLYNN, Patrick J. Data clustering: a review. **ACM computing surveys (CSUR)**, v. 31, n. 3, p. 264-323, 1999.

JAIN, Anil K.; DUIN, Robert P.. W.. ; MAO, Jianchang. Statistical pattern recognition: A review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4-37, 2000.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, n. 8, p. 651-666, 2010.

JOLLIFFE, I. T. Mathematical and statistical properties of population principal components. **Principal Component Analysis**, p. 10-28, 2002.

JURICEK, B. C.; SEBORG, D. E.; LARIMORE, W. E. Identification of the Tennessee Eastman challenge process with subspace methods. **Control Engineering Practice**, v. 9, n. 12, p. 1337–1351, dez. 2001.

KANTARDZIC, Mehmed. **Data mining: concepts, models, methods, and algorithms**. John Wiley & Sons, 2011.

KEOGH, Eamonn; LIN, Jessica. Clustering of time-series subsequences is meaningless: implications for previous and future research. **Knowledge and information systems**, v. 8, n. 2, p. 154-177, 2005.

LAU, C. K. et al. Fault diagnosis of Tennessee Eastman process with multi-scale PCA and ANFIS. **Chemometrics and Intelligent Laboratory Systems**, v. 120, p. 1-14, 2013.

LEE, Jong-Min *et al*. Nonlinear process monitoring using kernel principal component analysis. **Chemical Engineering Science**, v. 59, n. 1, p. 223-234, 2004.

LI, Chaoshun et al. A novel chaotic particle swarm optimization based fuzzy clustering algorithm. **Neurocomputing**, v. 83, p. 98-109, 2012.

LI, Gang et al. Generalized reconstruction-based contributions for output-relevant fault diagnosis with application to the Tennessee Eastman process. **IEEE Transactions on Control Systems Technology**, v. 19, n. 5, p. 1114-1127, 2011.

LI, Han; XIAO, De-yun. Fault diagnosis of Tennessee Eastman process using signal geometry matching technique. **EURASIP Journal on Advances in Signal Processing**, v. 2011, n. 1, p. 83, 2011.

LI, Lili; LIU, Xiyu; XU, Mingming. A novel fuzzy clustering based on particle swarm optimization. In: **Information Technologies and Applications in Education, 2007. ISITAE'07. First IEEE International Symposium on**. IEEE, 2007. p. 88-90.

LI, Shun; WEN, Jin. Application of pattern matching method for detecting faults in air handling unit system. **Automation in Construction**, v. 43, p. 49-58, 2014.

LIAO, T. W. *et al*. Understanding and projecting the battle state. In: **23rd Army Science Conference, Orlando, FL**. 2002.

LIAO, T. Warren. Clustering of time series data—a survey. **Pattern Recognition**, v. 38, n. 11, p. 1857-1874, 2005.

LIAO, T. Warren; TING, Chi-Fen; CHANG, Pei-Chann. An adaptive genetic clustering method for exploratory mining of feature vector and time series data. **International Journal of Production Research**, v. 44, n. 14, p. 2731-2748, 2006.

LIN, Yun-Ting; CHEN, Yen-Kuang; KUNG, Sun-Yuan. A principal component clustering approach to object-oriented motion segmentation and estimation. **The Journal of VLSI Signal Processing**, v. 17, n. 2, p. 163-187, 1997.

LIU, Jianzhuang; XIE, Weixin. A genetics-based approach to fuzzy clustering. In: **Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int**. IEEE, 1995. p. 2233-2240.

MACGREGOR, John; CINAR, Ali. Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods. **Computers & Chemical Engineering**, v. 47, p. 111-120, 2012.

MACQUEEN, James *et al*. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. 1967. p. 281-297.

MAGNUSSON, Willian E. **Estatística [sem] matemática: a ligação entre as questões ea análise**. Planta, 2003.

MAKI, Yunosuke; LOPARO, Kenneth A. A neural-network approach to fault detection and diagnosis in industrial processes. **IEEE Transactions on Control Systems Technology**, v. 5, n. 6, p. 529-541, 1997.

MANLY, Bryan FJ Manly. **Métodos estatísticos multivariados: uma introdução**. Bookman, 2008.

MANSFIELD, James R. *et al*. Fuzzy C-means clustering and principal component analysis of time series from near-infrared imaging of forearm ischemia. **Computerized Medical Imaging and Graphics**, v. 21, n. 5, p. 299-308, 1997.

MASUREL, Enno; NIJKAMP, Peter; VINDIGNI, Gabriella. Breeding places for ethnic entrepreneurs: a comparative marketing approach. **Entrepreneurship & Regional Development**, v. 16, n. 1, p. 77-86, 2004.

MAULIK, Ujjwal; BANDYOPADHYAY, Sanghamitra. Genetic algorithm-based clustering technique. **Pattern Recognition**, v. 33, n. 9, p. 1455-1465, 2000.

MAULIK, Ujjwal; BANDYOPADHYAY, Sanghamitra. Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. **IEEE Transactions on Geoscience and Remote Sensing**, v. 41, n. 5, p. 1075-1081, 2003.

MELGANI, Farid; BRUZZONE, Lorenzo. Classification of hyperspectral remote sensing images with support vector machines. **IEEE Transactions on Geoscience and Remote Sensing**, v. 42, n. 8, p. 1778-1790, 2004.

MELIN, Patricia; CASTILLO, Oscar. A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. **Applied Soft Computing**, v. 21, p. 568-577, 2014.

MENG, L.; WU, Q. H.; YONG, Z. Z. A genetic hard c-means clustering algorithm. **DYNAMICS OF CONTINUOUS DISCRETE AND IMPULSIVE SYSTEMS SERIES B**, v. 9, p. 421-438, 2002.

MICHALEWICZ, Z. Genetic Algorithms + Data Structures + Evolution Programs. Al Series, **Springer Verlag**, New York, 1994.

MICHIE, Donald; SPIEGELHALTER, David J.; TAYLOR, Charles C. **Machine learning, neural and statistical classification**. 1994.

MILLER, Harvey J.; HAN, Jiawei (Ed.). **Geographic data mining and knowledge discovery**. CRC Press, 2009.

MORO, Sergio; LAUREANO, Raul; CORTEZ, Paulo. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In: PROCEEDINGS OF EUROPEAN SIMULATION AND MODELLING CONFERENCE-ESM'2011. Hotel de Guimaraes, Guimaraes, Portugal (organized by University of Minho). **Anais**... Eurosis, 2011. p. 117-121.

MUKHOPADHYAY, Anirban; MAULIK, Ujjwal; BANDYOPADHYAY, Sanghamitra. Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. **IEEE Transactions on Evolutionary Computation**, v. 13, n. 5, p. 991-1005, 2009.

NAYAK, Janmenjoy; NAIK, Bighnaraj; BEHERA, H. S. Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In: **Computational Intelligence in Data Mining-Volume 2**. Springer, New Delhi, 2015. p. 133-149.

NOMA, Alexandre. **Duas abordagens para casamento de padrões de pontos usando relações espaciais e casamento entre grafos**. 2010. Tese de Doutorado. Universidade de São Paulo.

OLIVEIRA, H. *et al*. Inteligência Computacional aplicada a administração, economia e engenharia em Matlab. **Rio de Janeiro, Thompson**, 2007.

OLSHAUSEN, Bruno A.; ANDERSON, Charles H.; VAN ESSEN, David C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. **Journal of Neuroscience**, v. 13, n. 11, p. 4700-4719, 1993.

OLSZEWSKI, Robert T. **Generalized feature extraction for structural pattern recognition in time-series data**. CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2001.

PEDRYCZ, Witold. Fuzzy sets in pattern recognition: methodology and methods. **Pattern recognition**, v. 23, n. 1-2, p. 121-146, 1990.

RAHMAN, Md Anisur; ISLAM, Md Zahidul. A hybrid clustering technique combining a novel genetic algorithm with K-Means. **Knowledge-Based Systems**, v. 71, p. 345-365, 2014.

RANI, Sangeeta; SIKKA, Geeta. Recent techniques of clustering of time series data: a survey. **International Journal of Computer Applications**, v. 52, n. 15, 2012.

RATO, Tiago J.; REIS, Marco S. Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR). **Chemometrics and Intelligent Laboratory Systems**, v. 125, p. 101-108, 2013.

RIBEIRO, K. P; FONTES, C. H. O. A Genetic Algorithm Based Clustering Applied to Multivariate Time Series. In: 24TH ABCM INTERNATIONAL CONGRESS OF MECHANICAL ENGINEERING, 2017. Curitiba, PR, Brazil. **Anais**... ABCM, 2017.

RICARDEZ-SANDOVAL, L. A.; BUDMAN, H. M.; DOUGLAS, P. L. Simultaneous design and control of chemical processes with application to the Tennessee Eastman process. **Journal of Process Control**, v. 19, n. 8, p. 1377-1391, 2009.

RICKER, N. Lawrence. Decentralized control of the Tennessee Eastman challenge process. **Journal of Process Control**, v. 6, n. 4, p. 205-221, 1996.

ROKACH, Lior. A survey of clustering algorithms. In: **Data mining and knowledge discovery handbook**. Springer US, 2009. p. 269-298.

ROLLS-ROYCE. Material de Treinamento do Conjunto RB 211-G62 DF, 2010.

ROSÉN, Christian; YUAN, Z. Supervisory control of wastewater treatment plants by combining principal component analysis and fuzzy c-means clustering. **Water science and technology**, v. 43, n. 7, p. 147-156, 2001.

RUNKLER, Thomas A.; KATZ, Christina. Fuzzy clustering by particle swarm optimization. In: **Fuzzy Systems, 2006 IEEE International Conference on**. IEEE, 2006. p. 601-608.

SÁ BARRETO, S. T. **Desenvolvimento de Metodologia pata Atualização em Tempo Real de Modelos Matemáticos de Processos Decisórios**. Dissertação (Mestrado em Mecatrônica) – PPGM, Universidade Federal da Bahia, Salvador, 2009.

SÁEZ, Doris; CORTÉS, Cristián E.; NÚÑEZ, Alfredo. Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering. **Computers & Operations Research**, v. 35, n. 11, p. 3412-3438, 2008.

SALAHSHOOR, Karim; KORDESTANI, Mojtaba; KHOSHRO, Majid S. Fault detection and diagnosis of an industrial steam turbine using fusion of SVM (support vector machine) and ANFIS (adaptive neuro-fuzzy inference system) classifiers. **Energy**, v. 35, n. 12, p. 5472-5482, 2010.

SARAVANAMUTTOO, H. I. H.; ROGERS, G. F. C.; COHEN, H. **Gas Turbine Theory**. Dorchester: Prentice Hall, v. 5, 1996.

SCHKODA, Ryan F. **Clustering and Classification of Multivariate Stochastic Time Series in the Time and Frequency Domains**. 2012. Tese de Doutorado. Clemson University.

SHAMS, MA Bin; BUDMAN, H. M.; DUEVER, T. A. Fault detection, identification and diagnosis using CUSUM based PCA. **Chemical Engineering Science**, v. 66, n. 20, p. 4488-4498, 2011.

SINGHAL, Ashish; SEBORG, Dale E. Clustering multivariate time-series data. **Journal of chemometrics**, v. 19, n. 8, p. 427-438, 2005.

THOMAS, Michael C.; ZHU, Wenbo; ROMAGNOLI, Jose A. Data mining and clustering in chemical process databases for monitoring and knowledge discovery. **Journal of Process Control**, 2017.

TSENG, Vincent S. *et al*. Cluster-based genetic segmentation of time series with DWT. **Pattern Recognition Letters**, v. 30, n. 13, p. 1190-1197, 2009.

VARELLA, C. A. A. Análise de componentes principais.Universidade Federal Rural do Rio de Janeiro, Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf, 2008.

VENKATASUBRAMANIAN, Venkat *et al*. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. **Computers & Chemical Engineering**, v. 27, n. 3, p. 293-311, 2003.

WANG, Xiaozhe; SMITH, Kate; HYNDMAN, Rob. Characteristic-based clustering for time series data. **Data Mining and Knowledge Discovery**, v. 13, n. 3, p. 335-364, 2006.

WANG, Xiaoyue *et al*. Experimental comparison of representation methods and distance measures for time series data. **Data Mining and Knowledge Discovery**, p. 1-35, 2013.

WHITLEY, Darrell. A genetic algorithm tutorial. **Statistics and Computing**, v. 4, n. 2, p. 65-85, 1994.

WIKAISUKSAKUL, Siripen. A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering. **Applied Soft Computing**, v. 24, p. 679-691, 2014.

XU, Rui; WUNSCH, Donald. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, v. 16, n. 3, p. 645-678, 2005.

XUN, Lin; ZHISHU, Li. The similarity of multivariate time series and its application. In: MANAGEMENT OF E-COMMERCE AND E-GOVERNMENT (ICMECG), 2010 FOURTH INTERNATIONAL CONFERENCE ON. Chengdu, China. **Anais**… IEEE, 2010. p. 76-81.

YANG, M.-S. A survey of fuzzy clustering. Mathematical and Computer modelling, v. 18, n. 11, p. 1-16, 1993.

YAO, Yong; FREEMAN, Walter J. Model of biological pattern recognition with spatially chaotic dynamics. **Neural Networks**, v. 3, n. 2, p. 153-170, 1990.

YIN, Shen et al. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. **Journal of Process Control**, v. 22, n. 9, p. 1567-1581, 2012.

ZADEH, Lotfi A. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338-353, 1965.

ZHANG, Jun; CHUNG, Henry Shu-Hung; LO, Wai-Lun. Clustering-based adaptive crossover and mutation probabilities for genetic algorithms. **IEEE Transactions on Evolutionary Computation**, v. 11, n. 3, p. 326-335, 2007.

ZHANG, Junlin; WILLIAMS, Samuel Oluwarotimi; WANG, Haoxiang. Intelligent computing system based on pattern recognition and data mining algorithms. **Sustainable Computing: Informatics and Systems**, 2017.

UFBA
UNIVERSIDADE FEDERAL DA BAHIA
ESCOLA POLITÉCNICA

PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA INDUSTRIAL - PEI

Rua Aristides Novis, 02, 6º andar, Federação, Salvador BA
CEP: 40.210-630
Telefone: (71) 3283-9800
E-mail: pei@ufba.br
Home page: http://www.pei.ufba.br