IV SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R
R & PYTHON E AS TENDÊNCIAS DE COLABORAÇÃO
NITERÓI, 21 A 23 DE MAIO DE 2019

RECONHECIMENTO DE PADRÕES DE OPERAÇÕES DE PARTIDA DE TURBINA UTILIZANDO MACHINE LEARNING

Caroline Vasconcelos Fernandes[1]

Júlia Carolina Braz de Freitas Bijos[2]

Cristiano Hora de Oliveira Fontes[3]

Karla Patricia Santos Oliveira Rodriguez Esquerre[4]

Resumo

Diversos problemas interferem nos processos industriais e provocam impactos diretos na operação e nas estratégias a serem tomadas a longo prazo. Como consequência, há uma forte exigência por uma rápida resolução. O advento dos algoritmos de aprendizado de máquina possibilita que a atividade produtiva seja contemplada com análises de padrões e ferramentas estatísticas para tornar o processo mais eficiente. Nessa perspectiva, esse trabalho teve por objetivo verificar a aplicabilidade da técnica de agrupamento Fuzzy C-Means para séries temporais univariadas, utilizando diferentes métricas no reconhecimento de padrões no processo de operação de partida de turbinas. Para isso, foi aplicada a técnica de agrupamento de séries temporais Fuzzy C-Means, utilizando diferentes métricas de distância (Euclidiana e Dynamic Time Warping), com a utilização do Software R. A métrica utilizada não influenciou fortemente o comportamento das séries, porém, foi observado que a quantidade de grupos escolhidos pode comprometer a qualidade do agrupamento e a eficácia da técnica. Foi possível verificar o padrão existente quando a turbina apresenta falha, bem como o padrão de operação normal de partida. O Índice de Silhueta calculado na validação, indicou que para o estudo de caso da partida da turbina o número ideal é de dois grupos.
Palavras-chave: Agrupamentos, Fuzzy C-Means, Padrões , processos.

Abstract

[1] Programa de Pós-Graduação em Engenharia Industrial da Universidade Federal da Bahia (UFBA), carolinefernandes.eq@gmail.com.

[2] Programa de Pós-Graduação em Engenharia Industrial da Universidade Federal da Bahia (UFBA), juliabijos@outlook.com.

[3] Universidade Federal da Bahia (UFBA), cfontes@ufba.br

[4] Universidade Federal da Bahia (UFBA), karlaesquerre@ufba.br.

Many problems interfere on industrial processes causing direct impacts on operation and long term strategies. Hence, there's a strong request for a quick solution. The advent of machine learning algorithms enables the recognition of patterns and behaviors that aim to improve process efficiency. In this perspective, this paper aimed to verify applicability of the cluster technic Fuzzy C Means for univariate time series by using two different metrics of distance on pattern recognize on operation process of turbines start. The cluster method Fuzzy C Means was used by applying different distance metrics (Euclidiana and Dynamic Time Warping), utilizing Software R. The metrics used didn't influence the behaviour of time series, but, it was noted that the quantity of chosen groups may affect the quality of cluster and the technic efficiency. It was possible to note the existing pattern when turbine fails, and normal starting operation. The Silhouette index calculated on validation indicated that for this turbine start case study the ideal numbers of groups is two.
Keywords: Clusters, Fuzzy C-Means, patterns, processes.

Introdução

A necessidade de obtenção de dados, provenientes de processos industriais, é um desafio, visto que é preciso transformar esses dados em informações úteis, para auxiliar no mapeamento dos processos, aumentar a precisão na tomada de decisão e ainda, contribuir na definição das quantidades de produção. Assim, há um grande interesse na utilização de ferramentas que possibilitem a implementação eficiente de soluções, a partir das informações obtidas.

Para isso, técnicas de Aprendizado de Máquina (*Machine Learning*) tem sido extensivamente utilizadas no desenvolvimento de algoritmos que consigam extrair *insights* através do aprendizado das relações e tendências históricas nos dados. Tais técnicas estão divididas em duas abordagens: (i) aprendizado supervisionado, quando um conjunto de variáveis independentes é utilizado para prever uma variável dependente; e (ii) não-supervisionado, quando deseja-se obter representações significativas dos dados existentes, condensando a informação em pontos mais relevantes.

Dentre as técnicas existentes para aprendizado não-supervisionado, uma delas é a técnica de agrupamento (*clustering*) que tem como objetivo agrupar dados com comportamentos semelhantes em um mesmo grupo (*cluster*) e os distintos, em grupos diferentes.

Assim como é possível agrupar dados convencionais e inferir relações ou extrair padrões entre os mesmos, também é factível o agrupamento de séries temporais. O algoritmo de agrupamento Fuzzy C-Means (BEZDEK, 1981) tem sido uma das abordagens mais utilizadas para o agrupamento de séries temporais. É uma extensão do agrupamento clássico e tem como característica a associação de um grau de pertinência dos objetos a cada um dos grupos identificados/reconhecidos. Considerando os protótipos iniciais do algoritmo, o agrupamento do Fuzzy C-Means compreende a minimização da seguinte função objetivo:

$$J_q(U_h, C) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \, d(x_k, c_i) \quad (1)$$

Na equação 1, U_h é a matriz de pertinência com dimensão $c \times n$, onde c é a quantidade de grupos e n é o total de objetos na amostra. Desta forma, u_{ik} representa a pertinência do elemento k no *cluster* i, m é o índice de fuzzificação e $d(x_k, c_i)$ é a distância entre o cluster c_i e o objeto x_k.

Objetivo

O objetivo deste trabalho é realizar o reconhecimento de padrões do processo de operação de partida de turbinas, por meio de agrupamento Fuzzy C-Means utilizando diferentes métricas de similaridade.

Material e Método

O presente artigo utiliza dados de processos, relacionados à operação de uma turbina. Os dados utilizados correspondem respectivamente à 10 e 60 objetos de falha e de partida normal, de uma turbina a gás. Cada objeto compreende uma série temporal com 33 instantes de medição (aproximadamente 16 min) de temperatura de entrada do gás na turbina.

Na primeira etapa, foram retirados os rótulos (falha e operação normal) dos dados e em seguida normalizados, com base no mínimo e máximo da amostra. A normalização é importante para corrigir possíveis distorções (escala, translação, rotação, diferença de fase) que os dados brutos podem apresentar devido ao seu processo de aquisição. Logo após, gráficos normalizados foram gerados a fim de verificar o perfil das operações. Esta etapa é de fundamental importância visto que garante uma melhor visualização do perfil comportamental da série, além de permitir melhor avaliação, ao trazer os dados para a mesma escala de comparação, no presente caso, entre 0 e 1.

Posteriormente, foi aplicado o método de agrupamento Fuzzy C-Means (FCM), inicialmente utilizando a distância Euclidiana e variando o número de grupos, 2 e 3 grupos. Logo após, foi executada a técnica FCM com a métrica DTW.

Na última etapa, foi realizada a verificação e validação dos grupos obtidos, através do Índice Silhueta.

Optou-se por se trabalhar com o software R devido à melhor visualização gráfica e legibilidade do código, bem como a vantagem de ser uma linguagem de programação *open-access* e ser utilizada tanto no meio científico quanto no comercial. Desta forma, foram utilizados os seguintes pacotes disponíveis na biblioteca do software: ppclust, dplyr, ggplot2, reshape2, tidyr, dtwclust, factoextra.

Resultados e Discussão

Na Figura 1 e 2 é mostrado o comportamento da temperatura de entrada do gás da turbina após a normalização dos dados, para as séries temporais de falha e partida normal.

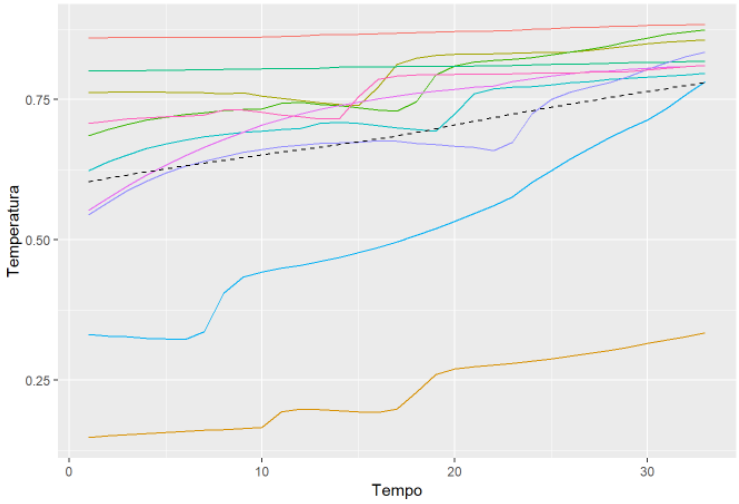

Figura 1 - Séries temporais de falha

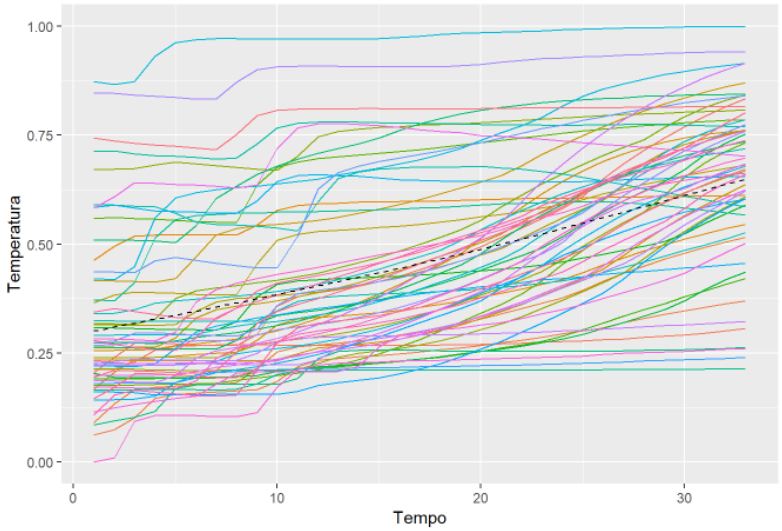

Figura 2 - Séries de temporais de partida normal

Tendo em vista o perfil apresentado nas Figuras 1 e 2, tem-se que os valores de temperatura se encontram normalizados numa escala entre 0 e 1. Assim, foi observado que a partida da turbina falha nos primeiros instantes em que a temperatura está elevada. Além disso, percebeu-se na Figura 1 que duas das séries temporais de falha se distanciam do comportamento das demais. Em relação a Figura 2, percebeu-se que a operação normal de uma turbina a gás, ocorre com mais chance em baixa-média temperatura.

Após a normalização dos dados das séries da turbina e posterior visualização, foi realizado o agrupamento através do método Fuzzy C-Means utilizando as métricas Euclidiana e DTW (*Dynamic Time Warp)* e variando o número de grupos. As Figuras 3 e 4 apresentam os resultados obtidos através do agrupamento Fuzzy C-means com distância Euclidiana, para dois e três grupos, respectivamente.

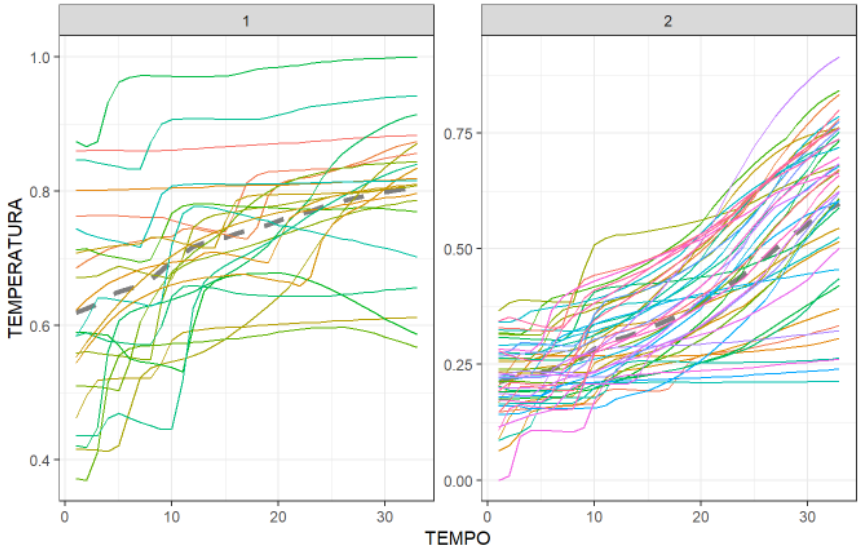

Figura 3 - Agrupamento FCM para 2 grupos, métrica Euclidiana.

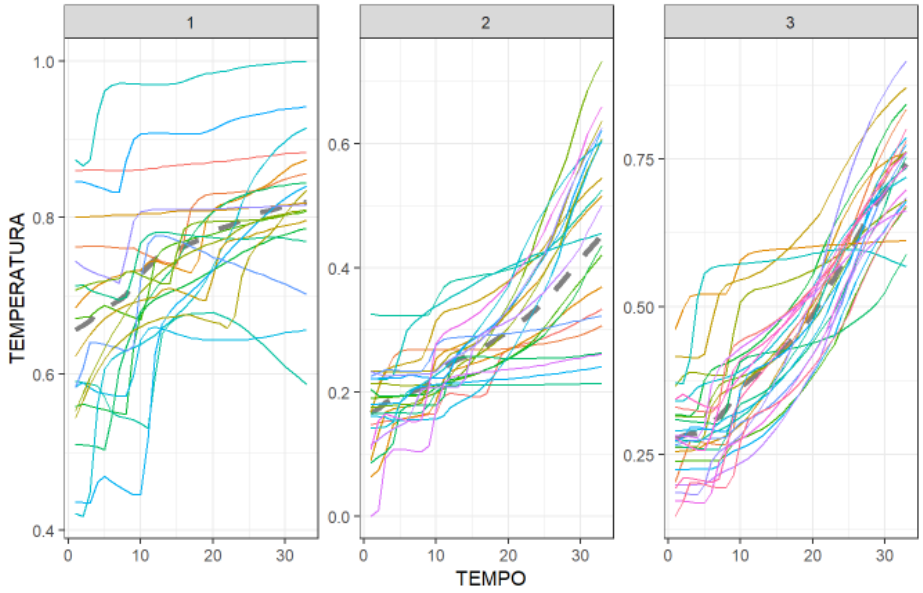

Figura 4 - Agrupamento FCM para 3 grupos, métrica Euclidiana

Tendo em vista o resultado apresentado na Figura 3, observou-se que a aplicação do FCM para dois conjuntos de séries, segue a tendência representada na figura 1 e 2, exceto o fato de que as duas únicas séries que iniciam em temperaturas mais baixas (Figura 1), não aparecem mais após a aplicação da técnica de agrupamento. Já na Figura 4, o agrupamento com 3 grupos gerou uma divisão da série de dados em mais de um grupo, no caso 2 e 3, enquanto que o grupo 1 permaneceu com o mesmo perfil. Nas Figuras 5 e 6 são apresentados

os agrupamentos FCM resultantes com a utilização da métrica DTW, com 2 e 3 grupos respectivamente.

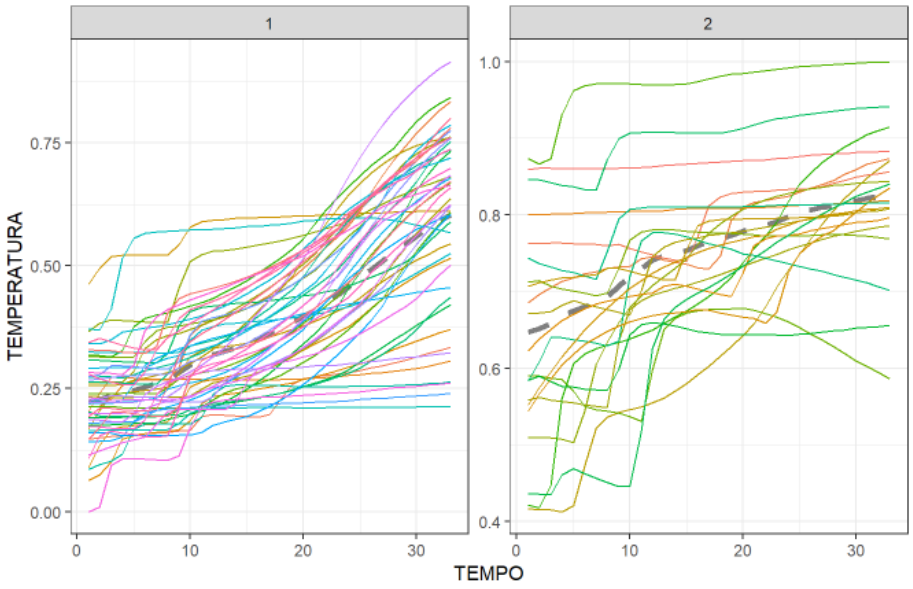

Figura 5 - Agrupamento FCM, para 2 grupos, métrica DTW

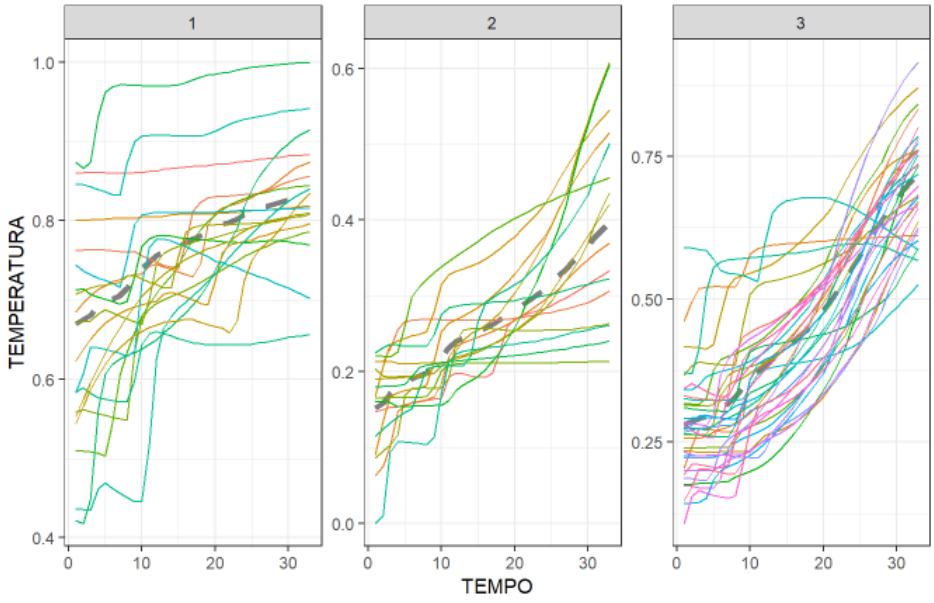

Figura 6 - Agrupamento FCM, 3 grupos, métrica DTW

Tendo em vista os resultados apresentados na Figura 5 e 6 foi possível notar que os resultados oriundos dos agrupamentos utilizando a métrica DTW, não apresentaram grandes

diferenças dos agrupamentos obtidos através da métrica euclidiana. Isto implica em dizer que para este caso, a modificação da métrica não exerceu grande influência no agrupamento final. Porém, sabe-se que a métrica DTW tem a vantagem de permitir alinhamentos não lineares em séries temporais, ou seja, a mesma permite o cálculo da distância a partir de dados que apresentam defasagem temporal.

Verificou-se que o grupo 1 da Figura 3, e grupo 2 da Figura 5, se assemelham ao perfil de falha apresentado inicialmente. Assim, inferiu-se que essa representação se caracterizava como um padrão de falha, caso a partida das turbinas fossem dadas. A partida de falha inicia com uma temperatura elevada em comparação com as partidas normais. Ou seja, o algoritmo de agrupamento proposto aplicado aos dados de processo das turbinas, conseguiu, de fato, captar o padrão de falha e partida normal da turbina.

Para a verificação e validação do agrupamento, utilizou-se o Índice de Silhueta, que define a qualidade do agrupamento com base na proximidade entre os objetos de um mesmo grupo e ao grupo mais próximo. As Figuras 7, 8 e 9 apresentam a aplicação do índice de Silhueta para diferentes números de grupos.

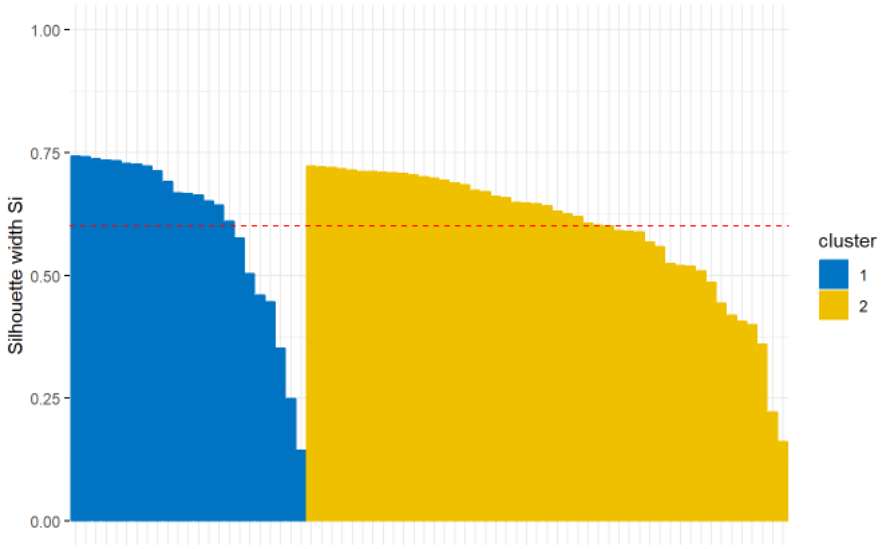

Figura 7 - Índice de Silhueta para 2 grupos. Valor médio 0,6

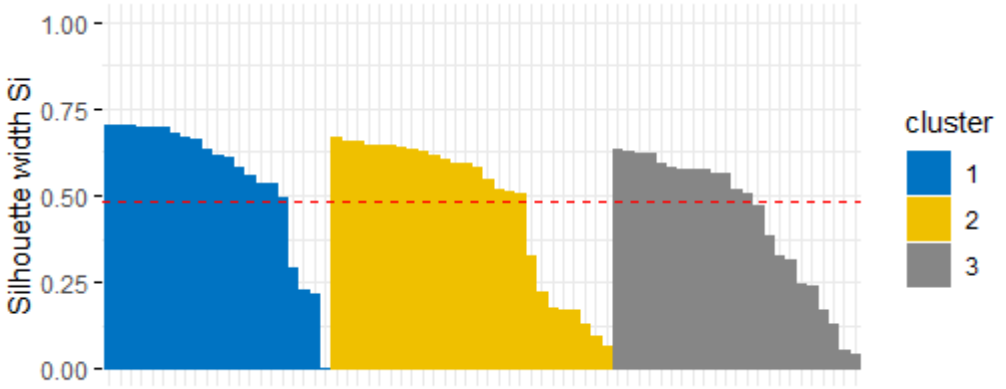

Figura 8 - Índice de Silhueta para 3 grupos. Valor médio 0,48
FONTE: Autores

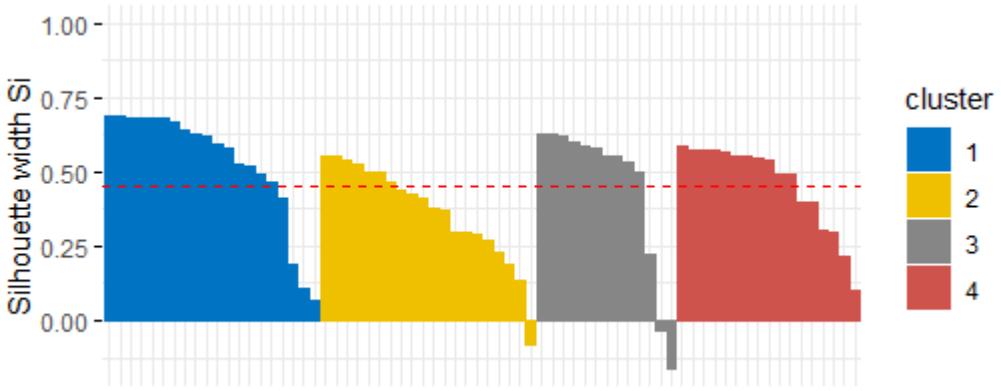

Figura 9 - Índice de Silhueta para 4 grupos. Valor médio 0,45
FONTE: Autores

Considerando os resultados de validação para os agrupamentos, das Figuras 7, 8 e 9, não foi verificado coalescência entre os grupos. Foram encontrados grupos bem separados e sem tendências a sobreposição. Sabe-se que quanto mais próximo de 1 o valor do índice, mais confiança se pode ter no agrupamento. Neste caso, observou-se que para 2 grupos o valor médio de silhueta, 0,6, indicou maior confiança no agrupamento, com maior homogeneidade dentro dos grupos.

Dessa forma foi possível perceber, que independentemente do tipo de métrica utilizada no agrupamento, o comportamento das séries não mostrou grandes mudanças. No entanto, o número de grupos escolhidos comprometeu a qualidade do agrupamento e consequentemente a eficácia da técnica, pois o índice silhueta mostra quão compactos e separados cada objeto se encontra dentro e fora de seu próprio grupo.

Conclusão

A utilização da técnica de agrupamento Fuzzy C-Means permitiu identificar o padrão existente quando a turbina apresenta falha relacionada às altas temperaturas de entrada do gás na partida, bem como o padrão de operação normal de partida. Isto torna esta técnica aplicável na identificação de falhas em processos industriais que apresentam características semelhantes com os dados utilizados neste trabalho. As métricas Euclidiana e DTW, não impactaram na tipologia das séries de falha e operação, entretanto, o desempenho da DTW tende a ser melhor, pois a mesma se mostra invariante à distorção de fase que as curvas podem apresentar entre si. A validação do agrupamento obtida através do Índice Silhueta, mostrou que o número de grupos sendo iguais a 2, ofereceu uma melhor qualidade de partição dos objetos, tanto no que se refere à compactação, quanto à separação entre os grupos.

O uso do R foi adequado para o processo de entendimento das características das séries em questão, possibilitando o uso das ferramentas necessárias para identificação dos padrões do estudo de caso analisado.

Referências

SOBRENOME, Nome. **Título:** subtítulo (se houver sem negrito). Edição (se houver). Local de publicação: Editora, data de publicação da obra

BEZDEK, J. C., 1981. **Pattern Recognition with Fuzzy Objective Function Algorithms.** New York: Plenum Press

COMPUTERWORLD.2013.**<https://computerworld.com.br/2013/03/20/big-data-o-desafio-de-garimpar-informacoes/>**. Acessado: 04 de Abril de 2019

CEBECI, Z., YILDIZ, F., KAVLAK, A.T., CEBECI, C. & Onder, H. (2018). **ppclust: Probabilistic and Possibilistic Cluster Analysis**. R package version 0.1.1, URL https://CRAN.R-project.org/package=ppclust

WICKHAM, H., FRANÇOIS, R., HENRY, L., MÜLLER, (2018). **dplyr: A Grammar of Data Manipulation**. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.

WICKHAM, H. 2007. **Reshaping Data with the reshape Package**. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

WICKHAM, H., HENRY, L. 2018. **tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions**. R package version 0.8.1. https://CRAN.R-project.org/package=tidyr

SARDA-ESPINOSA, A. 2018. **dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance**. R package version 5.5.1. https://CRAN.R-project.org/package=dtwclust.

KASSAMBARA, A., MUNDT, F. 2017. **factoextra: Extract and Visualize the Results of Multivariate Data Analyses**. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Anexo

Script

Aplicação da técnica de agrupamento Fuzzy C-Means à estudo de caso de operação de turbinas.

```
# Setando diretório de trabalho

setwd("C://Users/Usuario/Desktop/Ciência dos dados/AGRUPAMENTO/")

# Carregando bibliotecas necessárias

library(ppclust)

library(dplyr)

library(ggplot2)

library(reshape2)

library(tidyr)

library(dtwclust)

library(factoextra)

# Garantindo a reprodutibilidade

set.seed(42)

# Função para normalizar os dados utilizando MinMax

minmax <- function(x) {

    return((x - min(x)) / (max(x) - min(x)))

}

# Carregando os bancos de dados

df.falha <- read.csv("fault.csv")

df.normal <- read.csv("normal.csv")

df <- cbind(df.falha, df.normal)

# Normalizando os dados

df.normalizado <- minmax(df)

# Plotando séries temporais com falhas

df.falha.normalizado <- df.normalizado %>% select(1:10)

df.falha.normalizado <- df.falha.normalizado %>% mutate('time' = seq(1, 33, 1))
```

```
df.falha.normalizado <- df.falha.normalizado %>% gather(key ='objeto',value= 'medicao', 1:10)

df.falha.normalizado %>% ggplot(aes(x = time, y = medicao, color = objeto)) +

  theme(legend.position = "none") +

  labs(x = "Tempo", y = "Temperatura") + geom_line() + geom_smooth(method = "auto", se = FALSE, color = "black", size = 0.5, linetype = 2)

# Plotando séries temporais normais

df.normal.normalizado <- df.normalizado %>% select(11:70)

df.normal.normalizado <- df.normal.normalizado %>% mutate('time' = seq(1, 33, 1))

df.normal.normalizado <- df.normal.normalizado %>% gather(key = 'objeto', value = 'medicao', 1:60)

df.normal.normalizado %>% ggplot(aes(x = time, y = medicao, color = objeto)) +

   theme(legend.position = "none") +

   labs(x = "Tempo", y = "Temperatura") + geom_line() + geom_smooth(method = "auto", se = FALSE, color = "black", size = 0.5, linetype = 2)


# Plotando todas as séries temporais normalizadas

df.normalizado.fma <- df.normalizado %>% mutate('time' = seq(1, 33, 1))

df.normalizado.fma <- df.normalizado.fma %>% gather(key = 'objeto', value = 'medicao', 1:70)

df.normalizado.fma %>% ggplot(aes(x = time, y = medicao, color = objeto)) +

   theme(legend.position = "none") +

   labs(x = "Tempo", y = "Temperatura", subtitle = "SÃ©ries temporais") +

   geom_line() +

   geom_smooth(method = "auto", se = FALSE, color = "black", size = 0.5, linetype = 2)

# Transpondo os dados

df.normalizado.t <- t(df.normalizado)

##### Agrupamento FCM com disância Euclidiana
```

```
# Agrupando os dados com distância euclidiana e 2 clusters

grupos.2clusters.euc <- tsclust(series = df.normalizado.t, type = "fuzzy", k = 2L,

                                distance = "L2", centroid = "fcm")

p <- plot(grupos.2clusters.euc)

p + labs(x = "TEMPO", y = "TEMPERATURA", title = "Agrupamento FCM", subtitle =
"Métrica Euclidiana - cluster: 2") +

theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))


# Agrupando os dados com distância euclidiana e 3 clusters

grupos.3clusters.euc <- tsclust(series = df.normalizado.t, type = "fuzzy", k = 3L,

                                distance = "L2", centroid = "fcm")

p <- plot(grupos.3clusters.euc)

p + labs(x = "TEMPO", y = "TEMPERATURA", title = "Agrupamento FCM", subtitle =
"Métrica Euclidiana") +

   theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))


##### Agrupamento FCM com distância DTW


# Agrupando os dados com distância dtw e 2 clusters

grupos.2clusters.dtw <- tsclust(series = df.normalizado.t, type = "fuzzy", k = 2L,

                                distance = "dtw_basic", centroid = "fcm")

p <- plot(grupos.2clusters.dtw)

p + labs(x = "TEMPO", y = "TEMPERATURA", title = "Agrupamento FCM", subtitle =
"Métrica DTW") +

   theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))


# Agrupando os dados com distância dtw e 3 clusters
```

```
grupos.3clusters.dtw <- tsclust(series = df.normalizado.t, type = "fuzzy", k = 3L,
                                distance = "dtw_basic", centroid = "fcm")
p <- plot(grupos.3clusters.dtw)
p + labs(x = "TEMPO", y = "TEMPERATURA", title = "Agrupamento FCM", subtitle = "Métrica DTW") +
    theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))


####  Validação de grupos
# Índices
grupos.2centros <- eclust(df.normalizado.t, "fanny", hc_metric = "euclidean", k.max= 2)
fviz_silhouette(clus.dtw, palette = "jco",ggtheme = theme_minimal())
grupos.3centros <-eclust(df.normalizado.t, "fanny", hc_metric = "euclidean", k.max = 3)
fviz_silhouette(grupos.3centros, palette = "jco",ggtheme = theme_minimall())
grupos.4centros <-eclust(df.normalizado.t, "fanny", hc_metric = "euclidean", k.max= 4 )
fviz_silhouette(grupos.4centros, palette = "jco",ggtheme = theme_minimal())
```