

# An Interval Type-2 Maximum Likelihood Fuzzy Clustering Algorithm

Ben-Hur Matthews Moreno Montel<sup>1</sup>, Ginalber Luiz de Oliveira Serra<sup>2</sup>

<sup>1</sup>Programa de Pós-Graduação em Engenharia Elétrica – Universidade Federal do Maranhão (UFMA), São Luís – MA – Brazil

<sup>2</sup>Departamento de Eletroeletrônica (DEE) – Instituto Federal do Maranhão (IFMA) São Luís – MA – Brazil

ben-hur.matthews@discente.ufma.br, ginalber@ifma.edu.br

**Abstract.** *This paper presents a proposal for an algorithm for fuzzy clustering based on maximum likelihood processing over the data stream. The adopted methodology consists of eliminating initialization problems and mathematical determinations (convergence), related to the computational implementation, via the analysis of the distance norm between the data samples and the cluster centers, as well as the use of fuzzy type-2 interval systems to create realistic prototypes to clusters. Results related to benchmark data clustering and computational complexity analysis illustrate the efficiency of the proposed methodology compared to other clustering algorithms presented in the literature.*

**Resumo.** *Este artigo apresenta a proposta de um algoritmo para agrupamento nebuloso baseada no processamento de máxima verossimilhança sobre o fluxo de dados. A metodologia adotada consiste em eliminar problemas de inicialização e iderteminações matemáticas (convergência), relativas à implementação computacional, via análise da norma de distância entre as amostras dos dados e os centros dos agrupamentos, bem como no uso de sistemas nebulosos tipo-2 intervalares para a criação de protótipos realísticos aos agrupamentos. Resultados relacionados ao arupamento de dados benchmark e análise da velocidade de convergência ilustram a eficiência da metodologia proposta em comparação a outros algoritmos de agrupamento apresentados na literatura.*

## 1. Introdução

Os algoritmos de agrupamento são aplicados a classificação de dados em grupos de acordo com algum critério de similaridade, na qual uma análise revela os padrões escondidos existentes na base de dados. Estas classificações podem ser divididas em agrupamentos *hard* (cada ponto de dado pertence a um subgrupo) ou *soft* (cada ponto de dado pertence a um subgrupo com um valor de pertinência) [Babuška 1998, Pedrycz and Gomide 2007]. Para analisar as similaridades dentro de um conjunto de dados, é necessário minimizar uma função de custo baseada em uma norma de vetor. O *fuzzy C-means* (FCM) foi o primeiro algoritmo de agrupamento nebuloso proposto na literatura, e, após isso, várias adaptações foram desenvolvidas, criando uma série de algoritmos de agrupamento nebulosos [Ruspini et al. 2019]. Partições nebulosas permitem ao conjunto de dados possuir diferentes graus de pertinência relacionados a um agrupa-

mento, o que influencia no formato das partições. Nesse contexto, foram introduzidos os algoritmos de agrupamento nebuloso com matriz de covariância como mecanismo de adaptação à base de dados, onde os algoritmos Gustafson-Kessel (GK) e Estimação Fuzzy de Máxima Verossimilhança (EFMV) são bem conhecidos na literatura [Trauwaert et al. 1991, Gustafson and Kessel 1978]. O algoritmo de agrupamento EFMV apresenta problemas na etapa de inicialização devido à sua norma de distância exponencial, o que pode ser considerado como um fator limitante para a escolha desta técnica em aplicações práticas, pois esta gera problemas na convergência. As principais contribuições relacionadas à originalidade deste artigo são as seguintes:

- Uma versão nebulosa intervalar tipo-2 do algoritmo de agrupamento EFMV.
- Um nova abordagem que supera os problemas da inicialização e convergência do algoritmo de agrupamento EFMV.

## 2. Formulação da problemática

Algoritmos de agrupamento geram uma partição da base de dados, dividindo todas as informações em conjuntos menores com características semelhantes. O conjunto de dados é definido como  $n$  variáveis mensuráveis em  $N$  diferentes observações; assim tem-se que  $Z = [z_k | k = 1, 2, \dots, N]$ , onde  $z_k$  é um vetor coluna.

Em uma partição *hard*, a função objetivo gera valores unitários para alguns pontos na base de dados e nulo para outros. Em partições *soft*, cada ponto do conjunto de dados possui um grau de pertinência definido por  $\mu_{ik} \in [0, 1]$ . Nesse contexto, toda partição será considerada nebulosa, a qual é um tipo de partição *soft* [Babuška 1998, Pedrycz and Gomide 2007], desde que atenda às seguintes condições:

$$\mu_{ik} \in [0, 1] \quad (1)$$

$$\sum_{i=1}^c \mu_{ik} = 1, \quad 1 \leq k \leq N \quad (2)$$

$$0 < \sum_{k=1}^N \mu_{ik} < N \quad (3)$$

As similaridades são avaliadas pela distância entre um ponto ( $z_k$ ) e o protótipo do agrupamento ou centro ( $v_c$ ) do agrupamento. A partir de um conjunto de dados arbitrário ( $Z$ ), a tarefa de determinação da coleção de  $c$  agrupamentos é expressa pelo procedimento de minimização da seguinte função de custo [Pedrycz and Gomide 2007]:

$$J(Z; U, V) = \sum_{i=1}^c \sum_{k=1}^N (\mu_{ik})^m \|z_k - v_i\|_A^2 \quad (4)$$

onde o centro ou protótipo do agrupamento  $v_i$ , na qual  $m$  é o fator de ponderação fuzzy, é definido por:

$$v_i = \frac{\sum_{k=1}^N (\mu_{ik})^m z_k}{\sum_{k=1}^N (\mu_{ik})^m}, \quad 1 \leq i \leq c \quad (5)$$

A matriz de partição nebulosa  $U$  de dimensão  $c \times N$ , contém os graus de pertinência para cada dado em relação a cada agrupamento. Estes graus de pertinência são computados por:

$$\mu_{ik}^l = \frac{1}{\sum_{j=1}^c \left(\frac{D_{ikA}}{D_{jkA}}\right)^{\frac{2}{m-1}}} \quad (6)$$

Na Eq. (6) há dois termos a serem analisados, o primeiro é o fator de ponderação  $m \in [1, \infty]$ , que determina o quão nebulosa será a partição, e o segundo é norma de distância cujo valor é dado por:

$$D_{ikA}^2 = \|z_k - v_i\|_A^2 = (z_k - v_i)^T A (z_k - v_i) \quad (7)$$

onde  $A$  é uma matriz norma que determina o formato da partição, uma adaptação em  $A$  pode ser feita, de forma a seleciona-la como uma matriz de covariância, permitindo assim a mudança no padrão das partições de acordo com o conjunto de dados.

O algoritmo de agrupamento EFMV é baseado em uma matriz de covariância, enquanto sua norma de distância é uma função exponencial. Em consequência destes aspectos, cria-se a necessidade de uma boa inicialização devido à tendência do algoritmo convergir a um ótimo local [Babuška 1998, Trauwaert et al. 1991]. A norma de distância para o algoritmo de agrupamento EFMV é dada por:

$$D_{ik} = \frac{\det(F_i)^{0.5}}{P_i} \exp[0.5(z_k - v_i)^T F_i^{-1} (z_k - v_i)] \quad (8)$$

onde  $F_i$  é a matriz de covariância dos dados e  $P_i$  é a probabilidade a priori, computadas através das seguintes equações:

$$F_i = \frac{\sum_{k=1}^N (\mu_{ik})(z_k - v_i)(z_k - v_i)^T}{\sum_{k=1}^N (\mu_{ik})}, \quad 1 \leq i \leq c \quad (9)$$

$$P_i = \frac{1}{N} \sum_{k=1}^N \mu_{ik}, \quad 1 \leq i \leq c \quad (10)$$

Uma vez que a norma de distância é uma função exponencial, seu valor tende a crescer rapidamente. Toda linguagem de programação possui um valor numérico considerado como máximo, e sua rotina para lidar com operações que resultam em valores maiores que este valor máximo. No Matlab, após o valor máximo, todos os valores são considerados como infinito ( $\infty$ ). Por este motivo, há algumas operações como  $\infty \times 0$  e  $\frac{\infty}{\infty}$ , que não pode ser computadas. A função exponencial na norma de distância tende a resultar em números grandes para pontos de dados relativamente distantes dos centros dos agrupamentos, podendo violar o valor máximo suportado pela linguagem de programação, o que cria operações com indeterminação durante o processo de atualização da matriz de partição. Neste sentido, existem dois problemas: a inicialização do algoritmo de agrupamento EFMV que tende a aumentar o custo computacional da implementação, e a convergência devido à presença de indeterminações matemáticas.

## 2.1. Problemas de indeterminação matemática e inicialização do algoritmo EFMV

No algoritmo EFMV, é comum a aparição de indeterminações matemáticas, o que limita sua implementação. Para observar o surgimento e propagação destas indeterminações, considera-se que para um determinado agrupamento  $i$  e uma amostra  $z_k$  arbitrária, duas normas de distância tendem a infinito, conforme segue:

$$\begin{aligned}\lim_{z \rightarrow z_k} D_{i1k} &= \infty \\ \lim_{z \rightarrow z_k} D_{i2k} &= \infty\end{aligned}\quad (11)$$

Assim, ao expandir a equação da atualização da matriz de partição, obtêm-se:

$$\begin{aligned}\mu_{ik}(z_k) &= \frac{1}{\lim_{z \rightarrow z_k} \frac{D_{i1}}{x_1} + \dots + \lim_{z \rightarrow z_k} \frac{D_{i1}}{D_{i1}} + \dots + \lim_{z \rightarrow z_k} \frac{D_{i1}}{D_{i2}} + \dots + \lim_{z \rightarrow z_k} \frac{D_{i1}}{x_c}} \\ \mu_{ik}(z_k) &= \frac{1}{\infty + \dots + 1 + \dots + \infty + \dots + \infty} = NaN\end{aligned}\quad (12)$$

onde  $x_i$  é uma norma de distância com valor inferior a infinito. Como as equações (5-10) são funções de  $\sum_{k=1}^N \mu_{ik}$ , tal indeterminação é propagada para todas as variáveis. Assim, caso a norma de distância entre uma amostra e pelo menos dois protótipos tendam a infinito, o processo resultará em uma indeterminação matemática que, conseqüentemente, compromete a convergência do algoritmo EFMV.

Na literatura, é expresso que o algoritmo EFMV necessita de uma boa inicialização [Babuška 1998, Trauwaert et al. 1991]. Geralmente, o algoritmo EFMV é inicializado pela matriz de partição de outros algoritmos de agrupamento. Tal necessidade é um problema, pois além de aumentar o custo computacional, não há uma garantia de que tal inicialização impeça a aparição de indeterminações matemáticas. Neste contexto, este artigo propõe um algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar, de forma a superar possíveis problemas de inicialização e indeterminação matemática que comumente ocorrem no algoritmo tradicional EFMV.

## 3. Algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar

Um conjunto nebuloso tipo-1 apresenta uma incerteza determinística relacionada à função de pertinência. Entretanto, o conhecimento utilizado para construí-la é incerto, seja por subjetividade seja por ruído, o que torna necessário estender a incerteza para o formato das funções de pertinência [Karnik et al. 1999]. Um conjunto nebuloso tipo-2, por sua vez, aplica incerteza no formato das funções de pertinência ao permitir que cada ponto tenha grau de pertinência secundário. Entretanto, computar cada grau de pertinência é computacionalmente custoso. Ao assumir um conjunto intervalar, o custo computacional é reduzido, pois, agora, os graus de pertinência secundários são zero ou um, criando uma região de incerteza [Liang and Mendel 2000].

A importância dos sistemas nebulosos tipo-2 intervalares pode ser vista na variedade de trabalhos propostos na literatura [Gomes and Serra 2021, dos Santos Gomes and de Oliveira Serra 2021, dos Santos Gomes and de Oliveira Serra 2022, Karagöz et al. 2021, Precup et al. 2021]. Um conjunto nebuloso tipo-2 intervalar é definido por duas funções de pertinência, conhecidas como limites inferior e superior (subscrito + e - respectivamente). As operações

com conjuntos nebulosos tipo-2 intervalares são definidas em [Liang and Mendel 2000]. Para iniciar o algoritmo, a sua matriz de partição inicial  $U^0$  deve ser um conjunto nebuloso tipo-2 intervalar aleatório.

### 3.1. Evitando indeterminações

As indeterminações se apresentam no denominador de Eq. (12), no formato  $\frac{\infty}{\infty}$ . A ferramenta matemática para lidar com tal indeterminação é a regra de L'hospital, que para aplicá-la deve-se definir que as variáveis  $D_{ik}$ ...sejam funções de  $z_k$ . Assim, analisando os limites no denominador, obtém-se:

$$\lim_{z \rightarrow z_k} \frac{D_{i1}}{x_i} = \infty, \quad \lim_{z \rightarrow z_k} \frac{D_{i1}}{D_{i1}} = 1, \quad \lim_{z \rightarrow z_k} \frac{D_{i1}}{D_{i2}} = [0, \infty] \quad (13)$$

Como  $D_{ik}$  é uma norma de distância, seu valor é positivo. Assim, qualquer divisão entre as normas de distância é definida em  $[0, \infty]$ . Aplicando-se tais limites na EQ. (12) obtém-se:

$$\mu_{ik}(z_k) = \frac{1}{\infty + \dots + 1 + \dots + [0, \infty] + \dots + \infty} = 0 \quad (14)$$

Portanto, caso a norma de distância entre uma amostra  $z_k$  e um determinado protótipo  $v_i$  tenda a infinito, seu grau de pertinência para tal agrupamento é zero, como demonstrado em (14), evitando-se a aparição de indeterminações.

### 3.2. Algoritmo

Neste artigo, é proposto um algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar, inicializado por uma matriz de partição aleatória. A ideia principal deste algoritmo é ponderar a norma de distância exponencial da EFMV descrita na Eq. (8) com uma norma não exponencial no formato da Eq. (7), de maneira que em caso da condição de parada ( $CP = ||U^l - U^{l-1}||$ ) seja alta (próximo a 1) a norma não exponencial se sobressairá e, caso contrário, a norma de distância será influenciada de forma majoritária pela norma EFMV.

A fundamentação por trás desta proposta, é o fato de que os algoritmos com norma de distância não exponencial são inicializados com matriz de partição aleatória [Babuška 1998, Pedrycz and Gomide 2007, Trauwaert et al. 1991]. Assim, um  $CP$  alto indica que o algoritmo está longe de sua convergência, ou formato final, região onde a atuação da norma não exponencial é mais interessante. Já um  $CP$  próximo ao critério de tolerância indica que o formato da partição está próximo à convergência, podendo assim ser otimizado pela convergência próxima a um ótimo local da norma EFMV. O processo consiste em definir as zonas de operação exclusiva ( $O_{i1}$ ), onde qualquer  $CP$  superior a ela implicará em peso 1 na norma não exponencial ( $u_g = 1$ ) e qualquer  $CP$  inferior a  $O_{i2}$  implicará em peso 1 para a norma EFMV ( $u_m = 1$ ). Para qualquer valor fora das zonas de operação exclusivas, a ponderação será uma interpolação linear, de forma que os pesos são complementares.

O algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar é resumido nos seguintes passos:

1. Obtenha o conjunto de dados  $Z$ , escolha  $\tilde{m}$  e inicialize a matriz de partição intervalar aleatória  $\tilde{U}^0$ .

2. Calcule os protótipos ou centros dos agrupamentos  $\tilde{v}_c$  intervalares.

$$\tilde{v}_i^l = \frac{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)}) z_k}{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)})}, \quad 1 \leq i \leq c \quad (15)$$

3. Calcule a ponderação entre as normas de distância  $u_m$  e  $u_g$ .

4. Calcule as matrizes de covariância para a norma exponencial  $\tilde{F}_{MV}$  e não exponencial  $\tilde{F}_{NE}$ , e a probabilidade a priori  $\tilde{P}_i$ .

$$\tilde{F}_{MV}(i) = \frac{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)}) (z_k - \tilde{v}_i)(z_k - \tilde{v}_i)^T}{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)})}, \quad 1 \leq i \leq c \quad (16)$$

$$\tilde{F}_{NE}(i) = \frac{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)})^{\tilde{m}} (z_k - \tilde{v}_i)(z_k - \tilde{v}_i)^T}{\sum_{k=1}^N (\tilde{\mu}_{ik}^{(l-1)})^{\tilde{m}}}, \quad 1 \leq i \leq c \quad (17)$$

$$\tilde{P}_i = \frac{1}{N} \sum_{k=1}^N \tilde{\mu}_{ik}^{(l-1)} \quad 1 \leq i \leq c \quad (18)$$

5. Calcule as normas de distâncias intervalares necessárias, onde  $\rho$  é o volume do conjunto, geralmente é unitário.

$$\tilde{D}_{MV}(i, k) = \frac{\det(\tilde{F}_{MV}(i))^{0.5}}{\tilde{P}_i} \exp[0.5(z_k - \tilde{v}_i)^T \tilde{F}_{MV}(i)^{-1} (z_k - \tilde{v}_i)] \quad (19)$$

$$\tilde{D}_{NE}(i, k)^2 = (z_k - \tilde{v}_i)^T [\rho \det(\tilde{F}_{NE}(i))^{1/n} \tilde{F}_{NE}(i)^{-1}] (z_k - \tilde{v}_i) \quad (20)$$

$$\tilde{D}_{ik} = u_g \tilde{D}_{NE}(i, k) + u_m \tilde{D}_{MV}(i, k) \quad (21)$$

6. Atualize a matriz de partição intervalar.

Se  $D_{ik} > 0$  para  $1 \leq i \leq c, 1 \leq k \leq N$

$$\tilde{\mu}_{ik} = [\mu_{ik}^-, \mu_{ik}^+] \quad (22)$$

onde

$$\mu_{ik}^- = \min \left[ \frac{1}{\sum_{j=1}^c \frac{D_{jk}^-}{D_{jk}^-}}, \frac{1}{\sum_{j=1}^c \frac{D_{jk}^+}{D_{jk}^+}} \right] \quad (23)$$

$$\mu_{ik}^+ = \max \left[ \frac{1}{\sum_{j=1}^c \frac{D_{jk}^-}{D_{jk}^-}}, \frac{1}{\sum_{j=1}^c \frac{D_{jk}^+}{D_{jk}^+}} \right] \quad (24)$$

senão  $\tilde{\mu}_{ik} = [0, 0]$ .

Se  $D_{ik}^- \rightarrow \infty, u_{ik}^- \rightarrow 0$ , a mesma relação vale para o limite superior.

7. Repita do passo 2 ao 6 enquanto  $\|U^l - U^{l-1}\| > e$ , onde  $e$  é o critério de tolerância.

## 4. Resultados computacionais

Para validar a metodologia proposta, testes foram realizados com *benchmarks* existentes na literatura. O objetivo é avaliar o formato do agrupamento, o número de iterações até a convergência, o tempo decorrido durante o processo e a precisão da partição. As bases de dados selecionadas para os testes foram: modelo espiral, problema de Jain e IRIS. O modelo espiral e o problema de Jain, obtidas em [Fränti and Sieranoja 2018], contêm duas variáveis, o que facilita a observação do formato da partição. A IRIS, obtida em [Fisher 1936], possui quatro variáveis e, portanto, seus formatos não serão observados.

O *Adjusted Rand Index* (ARI) introduzido em [Anderson et al. 2010] é uma métrica para comparação entre partições *soft* e *hard*. Existem alguns trabalhos na literatura que avaliam o desempenho de diversos algoritmos de agrupamento utilizando o ARI [Zhou et al. 2022, Rathore et al. 2018], cujas comparações são feitas com uma matriz de partição de referência. Neste trabalho foi utilizada, como referência, a matriz de partição *hard* fornecida pela fonte dos dados. Ainda, neste artigo, o algoritmo proposto foi comparado aos algoritmos GK tipo-1, tipo-2 intervalar e o EFMV tipo-1. Para este último, foi utilizada a matriz de partição obtida no GK tipo-1 como inicialização e, além disso, um procedimento de correção para evitar indeterminações foi considerada. Os critérios de ponderação e convergência foram definidos como: expoente de ponderação nebulosa  $m = 2,1$  e  $\tilde{m} = [2, 1 \quad 2, 5]$ , zonas de operação  $O_{i1} = 0,45$  e  $O_{i2} = 0,15$ , e o critério de tolerância  $e = 0,0001$ .

As curvas de nível das partições são apresentadas nas figuras 1-6. As linhas representam níveis de pertinência, quanto mais próximo ao roxo, menor o grau de pertinência da região, quanto mais próximo ao amarelo, maior o grau de pertinência associado, tons de azul ou verde representam graus de pertinência intermediários.

### 4.1. Base de dados Jain

As figuras 1-4 apresentam o formato obtido pelos algoritmos GK-1, EFMV-1 e algoritmo proposto, considerando-se quatro clusters ( $c = 4$ ).

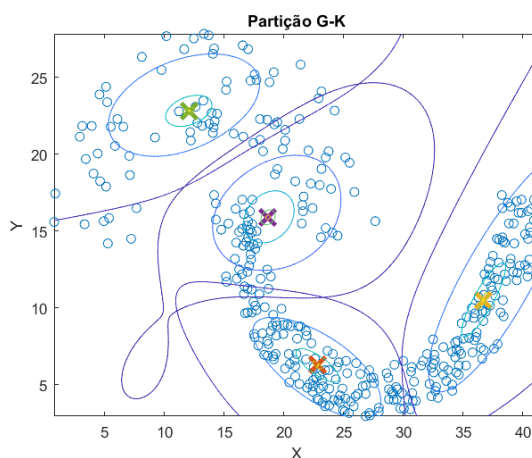
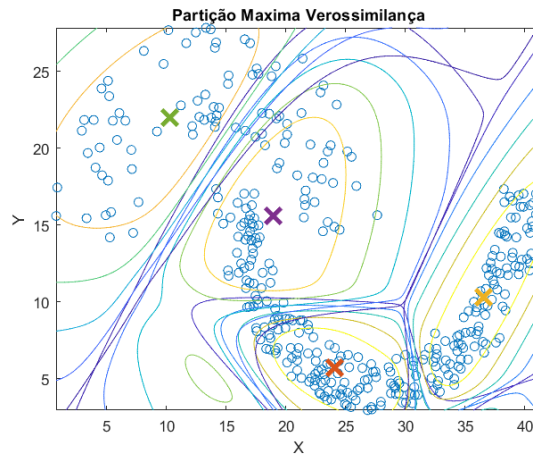
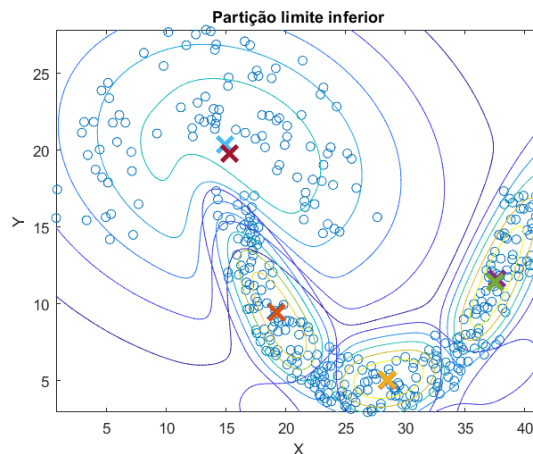


Figura 1. Curvas de nível da partição para p algoritmo de agrupamento fuzzy Gustafsson-Kessel tipo 1.



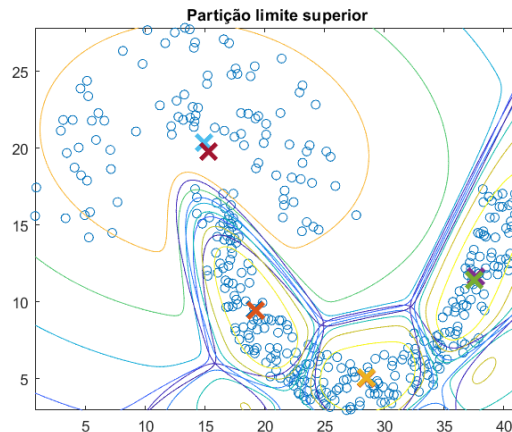
**Figura 2. Curvas de nível para a matriz de partição obtida no algoritmo EFMV tipo-1**



**Figura 3. Curvas de nível para a matriz de partição inferior obtida no algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar**

As figuras 1 e 2 apresentam os resultados para os algoritmos de agrupamento GK-1 e EFMV-1, respectivamente. Observa-se que tais figuras apresentam uma partição bem diferente do padrão apresentado pelos dados. A partição do algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar apresentou um padrão bem próximo à estrutura dos dados, mostrando-se visualmente superior para este conjunto de dados. Considerando-se que um dos objetivos dos algoritmos de agrupamento é dividir a base de dados em subconjuntos mais simples, o algoritmo de agrupamento proposto é superior, pois os formatos locais são mais simples, podendo-se observar características lineares em cada conjunto. Nas figuras 1 e 2 é notável a semelhança entre as curvas de nível. Isto mostra que além da redundância computacional, o formato da partição obtido pelo EFMV-1 é dependente da inicialização. Já no algoritmo de agrupamento proposto, a ponderação é governada pela necessidade da matriz de partição, resultando assim em uma formato mais fiel à base de dados.



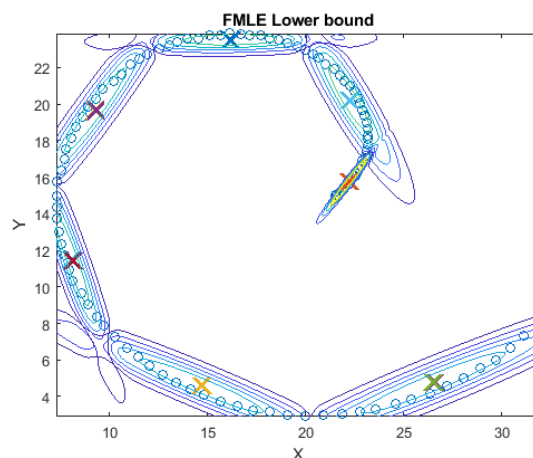


**Figura 4. Curvas de nível para a matriz de partição superior obtida no algoritmo nebuloso de máxima verossimilhança tipo-2 intervalar**

#### 4.2. Base de dados espiral

Para esta base de dados, por brevidade, apenas as curvas de nível do algoritmo proposto serão apresentadas, e os resultados comparativos com os outros algoritmos de agrupamento GK-1 e EFMV-1 serão analisados em termos de velocidade de convergência.

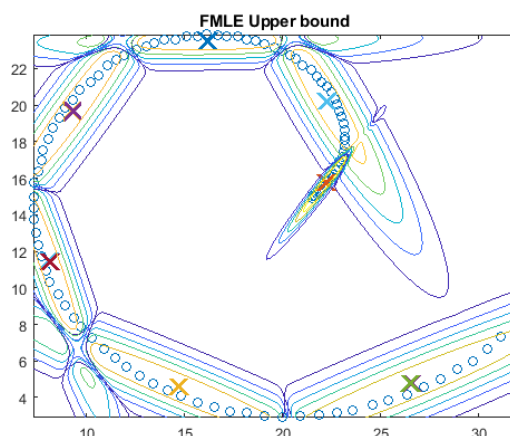
As figuras 5 e 6 apresentam as partições para os agrupamentos baseados no algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar. Pode-se observar que seus formatos são bem definidos, variam de acordo com a necessidade local dos dados, adicionando incerteza onde o conjunto de dados sugere, e acompanhando o fluxo de dados.



**Figura 5. Curva de nível para a matriz de partição inferior definida pelo algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar**

#### 4.3. Velocidade de convergência

Nesta seção, os algoritmos serão analisados em 3 diferentes aspectos: o número de iterações, o tempo decorrido durante o processo e o ARI. As tabelas 1- 3 apresentam



**Figura 6. Curva de nível para a matriz de partição superior definida pelo algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar**

tais resultados para as diferentes bases de dados testadas. As matrizes de partição obtidas pelos agrupamentos nebulosos foram comparadas com uma matriz de partição *hard* e, por isso, é esperado que o ARI não seja próximo a um. Nota-se, ainda, que os resultados do ARI não estão disponíveis para a base de dados espiral, pois não há matriz de partição para ser usada como referência (*soft* ou *hard*), de maneira que dificultou a comparação.

**Tabela 1. Análise da velocidade de convergência para a base de dados Jain.**

Algoritmo	Tempo (s)	Iterações	ARI
GK tipo-1	2,0344480 s	80	0.2328
EFMV tipo-1	3,69346 2s	94	0.2536
GK tipo-2	3,703922 s	100	[0.1748 0.2249]
Proposto	5,007393 s	148	[0.2466 0.3032]

**Tabela 2. Análise da velocidade de convergência para a base de dados IRIS.**

Algoritmo	Tempo (s)	Iterações	ARI
GK tipo-1	1.315 s	108	0.4507
EFMV tipo-1	2.7335 s	227	0.5922
GK tipo-2	1.6390 s	65	[0.2822 0.3427]
Proposto	2.66 s	88	[0.4698 0.6393]

A partir dos dados observados nas tabelas acima, pode-se constatar que o algoritmo proposto possui, em geral, a convergência mais lenta, porém não muito superior à de seus concorrentes. No quesito número de iterações, o algoritmo proposto apresenta um número compatível (inferior, em alguns casos) à seus concorrentes. Portanto, uma lógica de programação diferente nos passos computacionais do algoritmo pode reduzir consideravelmente o tempo necessário para sua convergência. Analisando-se a precisão pelo ARI, observa-se que o algoritmo proposto possui a melhor faixa de resultados entre os algoritmos testados. Assim o algoritmo proposto converge em um número de iterações

**Tabela 3. Análise da velocidade de convergência para a base de dados Espiral.**

Algoritmo	Tempo (s)	Iterações
GK tipo-1	3,63441 s	213
EFMV tipo-1	4,979702 s	304
GK tipo-2	2.739 s	120
Proposto	5,438501 s	209

compatível aos algoritmos comparados, porém sua iteração pode chegar a ser mais custosa, computacionalmente, com a garantia de uma partição mais precisa.

## 5. Comentários finais

O objetivo deste artigo é propor um algoritmo de agrupamento nebuloso de máxima verossimilhança tipo-2 intervalar, de forma a evitar a necessidade de uma inicialização e indeterminações matemática durante o processamento dos dados. Três bases de dados foram testadas e seus resultados foram analisados em termos do formato da partição, velocidade de convergência e precisão. No quesito formato da partição, observou-se a superioridade do algoritmo proposto, pois a ponderação entre as normas de distância fez com que o formato da matriz de partição se adaptasse segundo à necessidade da base de dados. No quesito velocidade de convergência, o algoritmo proposto apresentou uma iteração média mais custosa, devido à computação de duas normas de distância e matrizes de probabilidade, o que resultou em um tempo maior de convergência para certos casos. O algoritmo proposto apresentou a maior faixa de precisão em relação aos algoritmos comparados e, devido às características dos conjuntos nebulosos tipo-2 intervalares, adaptou-se melhor ao fluxo de dados, pois suas funções de pertinência apresentaram uma faixa (intervalo) de valores e não um valor específico. Como objeto de pesquisas futuras, considera-se a implementação da metodologia proposta no contexto nebuloso evolutivo como de particular interesse.

## 6. Referências

### Referências

- Anderson, D. T., Bezdek, J. C., Popescu, M., and Keller, J. M. (2010). Comparing fuzzy, probabilistic, and possibilistic partitions. *IEEE Transactions on Fuzzy Systems*, 18(5):906–918.
- Babuška, R. (1998). *Fuzzy modeling for control*. 1st edition.
- dos Santos Gomes, D. C. and de Oliveira Serra, G. L. (2021). Computational approach for real-time interval type-2 fuzzy kalman filtering and forecasting via unobservable spectral components of experimental data. 32:326–355.
- dos Santos Gomes, D. C. and de Oliveira Serra, G. L. (2022). Interval type-2 fuzzy computational model for real time kalman filtering and forecasting of the dynamic spreading behavior of novel coronavirus 2019. *ISA Transactions*, 124:57–68.
- Fisher, R. (1936). Iris data set. [urlhttps://archive.ics.uci.edu/ml/datasets/iris](https://archive.ics.uci.edu/ml/datasets/iris).
- Fränti, P. and Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, 48:4743–4759.

- Gomes, D. C. d. S. and Serra, G. L. d. O. (2021). Machine learning model for computational tracking and forecasting the covid-19 dynamic propagation. *IEEE Journal of Biomedical and Health Informatics*, 25(3):615–622.
- Gustafson, D. E. and Kessel, W. C. (1978). Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, pages 761–766.
- Karagöz, S., Deveci, M., Simic, V., and Aydin, N. (2021). Interval type-2 fuzzy aras method for recycling facility location problems. *Applied Soft Computing*, 102:107107.
- Karnik, N., Mendel, J., and Liang, Q. (1999). Type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, 7(6):643–658.
- Liang, Q. and Mendel, J. (2000). Interval type-2 fuzzy logic systems: theory and design. *IEEE Transactions on Fuzzy Systems*, 8(5):535–550.
- Pedrycz, W. and Gomide, F. (2007). *Fuzzy Systems Engineering Toward Human-Centric Computing*. Jhon Wiley and Sons.
- Precup, R.-E., David, R.-C., Roman, R.-C., Szedlak-Stinean, A.-I., and Petriu, E. M. (2021). Optimal tuning of interval type-2 fuzzy controllers for nonlinear servo systems using slime mould algorithm. *International Journal of Systems Science*, 0(0):1–16.
- Rathore, P., Bezdek, J. C., Erfani, S. M., Rajasegarar, S., and Palaniswami, M. (2018). Ensemble fuzzy clustering using cumulative aggregation on random projections. *IEEE Transactions on Fuzzy Systems*, 26(3):1510–1524.
- Ruspini, E. H., Bezdek, J. C., and Keller, J. M. (2019). Fuzzy clustering: A historical perspective. volume 14, pages 45–55.
- Trauwert, E., Kaufman, L., and Rousseeuw, P. (1991). Fuzzy clustering algorithms based on the maximum likelihood principle. volume 42, pages 213–227.
- Zhou, J., Pedrycz, W., Gao, C., Lai, Z., Wan, J., and Ming, Z. (2022). Robust jointly sparse fuzzy clustering with neighborhood structure preservation. *IEEE Transactions on Fuzzy Systems*, 30(4):1073–1087.