Avaliação de método de mensuração da qualidade da rede neural SOM para a tarefa de clusterização

Vinícius Leite Xavier[1], **Marcos Aurélio Santos da Silva**[2]

[1]Bolsista PIBITI/Embrapa/CNPq, Departamento de Computação
Universidade Federal de Sergipe, São Cristóvão, Sergipe

[2]Embrapa Tabuleiros Costeiros
Av. Gov. Paulo Barreto de Menezes, 3250, 49025-040, Aracaju, Sergipe

vinicius.xavier@dcomp.ufs.br, marcos.santos-silva@embrapa.br

Abstract. *This work evaluates a method for selecting Self-Organizing Maps based on the quality of topological preservation and representation of the neural network input data for clustering benchmark data based on the segmentation of the neural network weights. We evaluated five clustering algorithms: k-means, hierarchical agglomerative, and three methods based on graph partitioning. The results showed that the method for selecting the best neural network was effective for all four databases evaluated, although it did not generate optimal results. We observed that the performance of the clustering algorithms varies according to the type of data, with k-means presenting good performance for hyperspherical data and for the Iris database, the agglomerative hierarchical method being more effective for the MNIST database, and a method based on graph partitioning being more effective for data with arbitrary structure.*

Resumo. *Este trabalho avalia um método de seleção de Mapas Auto-Organizados a partir da qualidade da preservação topológica e representação dos dados de entrada da rede neural para fins de clusterização de dados de benchmark a partir da segmentação dos pesos da rede neural. Foram avaliados cinco algoritmos de clusterização: k-médias, hierárquico aglomerativo e três métodos baseados no particionamento de grafos. Os resultados mostraram que o método de seleção da melhor rede neural foi efetivo para todas as quatro bases de dados avaliadas, embora não tenha gerado resultados ótimos. Observou-se que o desempenho dos algoritmos de clusterização varia conforme o tipo de dado, com o k-médias apresentando bom desempenho para dados hiperesféricos e para a base Iris, o método hierárquico aglomerativo sendo mais efetivo para a base MNIST e um método baseado no particionamento de grafos mais efetivo para dados com estrutura arbitrária.*

1. Introdução

Os Mapas Auto-Organizáveis (SOM) são um tipo de rede neural artificial não-supervisionada, com aprendizado competitivo. Destaca-se na visualização, compressão e, principalmente, na clusterização de dados [Kohonen 2001]. Essa rede neural consegue reduzir grandes volumes de dados de entrada n para um número muito menor m de vetores de código, ou pesos, mantendo as características estatísticas e topológicas dos dados originais. Essa capacidade permite que a clusterização seja realizada nos pesos da rede

em vez dos dados, utilizando algoritmos clássicos como k-médias ou heurísticas baseadas em grafos, como as de [Silva et al. 2024, Silva and Costa 2011, Costa and Netto 2003]. A grande vantagem do SOM reside em sua habilidade de representar estruturas de dados complexas, que seriam difíceis de identificar por outros métodos, como exemplificado em estudos sobre a diversidade da agropecuária brasileira [Silva et al. 2022] e o mapeamento de culturas com dados satelitais [Santos et al. 2021].

Apesar do bom desempenho da rede SOM nessas tarefas devido a sua robustez, um desafio se impõe. A definição dos seus hiperparâmetros como dimensão (1D ou 2D), número de neurônios (vetores de código), função de vizinhança, formato da grade (retangular, hexagonal) e taxa de aprendizagem. Além disso, temos que para o processo de aprendizado sequencial estocástico e não determinístico, a rede SOM pode gerar resultados diferentes para a mesma parametrização inicial. A maioria dos trabalhos apresentados na literatura utilizam o erro de quantização ou o erro topológico como indicador da qualidade da representação dos dados pelos pesos da rede SOM [Appukuttan et al. 2025, Hameed et al. 2024]. No entanto, como observou [Delgado et al. 2017] essas medidas não avaliam a qualidade da representação topológica dos dados de entrada pelos vetores de código. Ou seja, os pesos podem se aproximar dos dados de entrada, mas não representarem corretamente sua topologia.

Dos índices propostos na literatura para avaliar a qualidade da rede neural SOM treinada quanto ao mapeamento topológico destacamos o Erro Combinado [Kaski and Lagus 1996] e a função topográfica [Villmann et al. 1997]. No primeiro temos a combinação do erro de quantização com o erro topográfico representado pela distância na grade neural entre o primeiro e o segundo *Best Match Unit* (BMU) para cada vetor de entrada. A função topográfica se apoia na definição de campos receptivos para a determinação da preservação topológica. Nesse caso, a rede neural terá preservado a topologia se dados adjacentes no espaço de entrada forem mapeados a neurônios vizinhos, e se neurônios adjacentes na grade neural forem mapeados a campos receptivos vizinhos no espaço de entrada.

A partir das características desses indicadores de qualidade da rede SOM, Delgado et al. (2017) propõem um método de escolha da melhor rede SOM. Primeiro as redes SOM seriam agrupadas por grupos com diferentes hiperparâmetros exceto o número de neurônios. De cada grupo desses seria extraída a rede SOM com menor Erro cominado. E dentre essas redes selecionadas com diferentes tamanhos seria escolhida aquela com menor valor para a função topográfica, já que neste caso não teríamos o efeito do tamanho da rede sobre esse indicador.

Para a clusterização da rede SOM observa-se pelo menos três abordagens. Na primeira, é definida uma rede SOM unidimensional onde cada nerônio representa um grupo [Delgado et al. 2017]. Na segunda é aplicado um método de clusterização sobre os pesos da rede neural, em geral k-médias ou hierárquico aglomerativo [Silva et al. 2022, Santos et al. 2021]. Na terceira abordagem, os pesos são segmentados a partir de uma estratégica de particionamento de grafos [Silva et al. 2024, Silva and Costa 2011, Costa and Netto 2003]. Neste último caso são usadas informações da rede SOM como número de observações associadas a cada neurônio, densidade de observações entre neurônios, distância entre os vetores de peso etc. Segundo [Delgado et al. 2017] e [Melo Riveros et al. 2019] o método k-médias tende a encontrar grupos com número ba-

lanceados de elementos e é mais susceptível a mínimos locais, enquanto que no método hierárquico aglomerativo seria difícil definir uma heurística para particionamento do dendograma [Vesanto and Alhoniemi 2000].

Este estudo parte da hipótese que o desempenho do algoritmo de clusterização está relacionado às características dos dados. Desta forma propomos avaliar o desempenho de diferentes métodos de clusterização de dados baseados no Mapa Auto-Organizável de Kohonen sobre quatro conjuntos de dados de *benchmark*. No trabalho, avaliamos o método proposto por [Delgado et al. 2017] para seleção da melhor rede SOM para a tarefa de clusterização de diferentes conjuntos de dados a partir da segmentação dos pesos W da rede SOM treinada usando cinco algoritmos: k-médias, hierárquico aglomerativo, e três baseados no particionamento de grafos [Silva et al. 2024, Silva and Costa 2011, Costa and Netto 2003]. Foram avaliadas as correlações entre os índices de qualidade de preservação topológica (erro combinado e função topográfica) e o tamanho m da rede neural usada na clusterização.

2. Materiais e métodos

2.1. Mapas Auto-Organizáveis

Dado um conjunto de dados de dados $\boldsymbol{X}$ no espaço $M \subseteq \mathbb{R}^d$ podemos construir uma rede neural não-supervisionada Mapa Auto-Organizável com grade no espaço A do espaço d-dimensional com m neurônios e topologia retangular ou hexagonal, podemos gerar um mapeamento com preservação topológica $\mathcal{M}_A$ do dado no espaço A. Para cada neurônio $i \in A$ temos um vetor de pesos $w_i \in \Re^d$ associado. O mapeamento $\mathcal{M}_A = (\Psi_{A\to M}, \Psi_{M\to A})$ de M para A é definido por $\Psi_{M\to A}$ e o inverso por $\Psi_{A\to M}$.

$$\mathcal{M}_A = \begin{cases} \Psi_{M\to A} : M \to A; & x \in M \mapsto i^*(x) \in A \\ \Psi_{A\to M} : A \to M; & i \in A \mapsto w_i \in M \end{cases} \tag{1}$$

com $i^*(x)$ como representação do neurônio i e seu peso $w_{i^*(x)}$ mais próximo de v (*Best Match Unit*), com $\left\|w_{i^*(x)} - v\right\| \leq \|w_j - x\|$. A rede neural SOM proposta por Kohonen (2001) é uma rede bidimensional artificial que possui neurônios representados por vetores de pesos distribuídos em uma grade retângular ou hexagonal. O processo de aprendizagem de máquina dessa proposta de rede é divido em três etapas: competitiva, cooperativa e adaptativa. Na fase competitiva, os valores de entrada de dados são apresentados à rede neural, a qual seleciona o vetor de pesos da grade com a menor distância ao vetor de entrada, neurônio $i^*(x)$. Logo em seguida, na fase cooperativa, a vizinhança é então definida com base na função de vizinhança $h(t)$ (e.g., função Gaussiana). Na fase adaptativa os valores dos pesos dos neurônios associados à rede na fase adaptativa são atualizados de acordo com $w(t+1) = w(t) + \alpha(t)h(t)(x(t) - w(t))$, que representa a atualização dos pesos W no tempo t e com função da taxa de aprendizagem $\alpha(t)$.

2.2. Avaliação da qualidade da rede SOM

Após o processo de aprendizado de máquina e obtenção de uma rede SOM treinada que possui cada neurônio associado a uma observação da base de dados, pode-se, segundo [Kohonen 2001], aferir a qualidade dessa rede a partir da avaliação da representatividade

dos dados X pelos pesos W e pela qualidade da preservação topológica do mapeamento $\mathcal{M}_A$ por métodos, como o erro de quantização e a função topográfica.

O erro de quantização $E_q = \frac{\sum_{k=1}^{n} \left\| x_k - w_{i^*(x)} \right\|}{n}$ representa a média das distâncias entre cada amostra de dados ao vetor de pesos do neurônio de melhor correspondência (BMU) mas não avalia a preservação topológica [Kohonen 2001]. O Erro topológico corresponde à proporção de observações cujo segundo BMU não é vizinho, na grade neural, do primeiro BMU. Neste caso, temos uma avaliação indireta e incompleta da preservação topológica.

Dado A, uma grade retangular de dimensão d_A, e M, um manifold $M \subseteq \Re^d$. Um mapa $\mathcal{M}_A = (\Psi_{A \to M}, \Psi_{M \to A})$ de M é preservado topologicamente se ambos os mapeamentos $(\Psi_{M \to A})$ de M para A e $(\Psi_{A \to M})$ de A para M são preservados em relação à vizinhança.

- O mapeamento $(\Psi_{M \to A})$ é preservado em relação à vizinhança se somente se a localização dos vetores w_i, w_j que são adjacentes em M pertencem a vértices i, j que são também adjacentes em A, segundo a norma máxima $||.||_{max}$.

- O mapeamento $(\Psi_{A \to M})$ é preservado em relação à vizinhança se somente se a localização dos vértices i, j que são adjacentes em A, de acordo com a norma euclidiana ou de acordo com a soma das normas $||.||_{\sum}$, estão relacionados a vetores de pesos w_i ,$w_j \in M$ vizinhos.

Destacamos duas métricas que levam em consideração a preservação topológica da rede neural SOM: a função topográfica [Villmann et al. 1997] e o erro combinado [Kaski and Lagus 1996], adotadas na proposta de [Delgado et al. 2017] para escolha da melhor rede SOM treinada.

2.2.1. Função topográfica

Para a grade de neurônios A, calcula-se a triangulação induzida de Delaunay D_M, grafo que conecta apenas os vetores de pesos w_i e w_j com regiões adjacentes no poliedro mascarado de Voronoi $\tilde{V}_i$, $\tilde{V}_j$, sendo que $d_{D_M}(i,j)$ representa a métrica associada à menor distância entre dois neurônios i, j em D_M e $\#\{.\}$ denota a cardinalidade do conjunto. Define-se as funções de preservação topológica $(\Psi_{M \to A})$ e $(\Psi_{A \to M})$ conforme as Eqs. $f_i(k) \stackrel{def}{=} \#\{j| \ ||\, i-j \,||_{max} > k; d_{D_M}(i,j) = 1\}$ e $f_i(-k) \stackrel{def}{=} \#\{j| \ ||\, i-j \,||_E = 1; d_{D_M}(i,j) > k\}$.

Desse modo, a função $f_i(k)$ mede preservação da vizinhança do mapeamento de M em A ao computar casos em que neurônios não vizinhos em A possuem uma relação de vizinhança em D_M segundo a métrica $d_{D_M}(i,j)$, enquanto a função $f_i(-k)$ mede a preservação da vizinhança de A em M, ao contabilizar casos de neurônios vizinhos em A que não sejam vizinhos em D_M. Segundo [Villmann et al. 1997], a função topográfica ϕ_A^M pode então ser definida como a média dos valores da função f_i para todos componentes j da grade de neurônios A, onde $k = 1, ..., m-1$. Conforme descrito pela Eq. 2.

$$\phi_A^M(k) \stackrel{def}{=} \begin{cases} \frac{1}{N}\sum_{j\in A} f_j(k) & k > 0 \\ \phi_A^M(1) + \phi_A^M(-1) & k = 0 \\ \frac{1}{N}\sum_{j\in A} f_j(k) & k < 0 \end{cases} \quad (2)$$

Conforme [Delgado et al. 2017], o valor de ϕ_A^M para $k = 0$ representa a combinação das medidas de não conformidades da preservação topológica $f_i(k)$ e $f_i(-k)$, sendo portanto, usada como medida de referência da função topográfica. Quanto menor $\phi_A^M(0)$ melhor a rede SOM preservou a topologia dos dados de entrada. Neste caso, à medida que a rede SOM cresce a primeira medida tende a crescer, enquanto a segunda tende a ser menor. Dessa forma, segundo [Delgado et al. 2017], uma medida anularia a outra de forma que para a função topográfica não haveria correlação entre o número de neurônios e o somatório destas duas medidas.

2.2.2. Erro cominado (EC)

Para uma dada amostra de dados x_i, primeiro calculamos suas duas melhores unidades correspondentes (BMUs), i_1^* e i_2^*. Em seguida, calculamos a soma das distâncias euclidianas de x_i até o vetor protótipo $w_{i_2^*}$ do segundo BMU, começando com a distância de x_i a $w_{b_1^*}$ e, posteriormente, seguindo o caminho mais curto até $w_{b_2^*}$, passando apenas pelas unidades vizinhas no mapa. Seja p um caminho no mapa de comprimento $P >= 1$, de $p(0) = i_1^*$ a $p(L) = i_2^*$, tal que $p(k)$ e $p(k+1)$ devem ser vizinhos para $k = 0...P-1$. A distância ao longo do caminho mais curto no mapa é calculada como: $EC_i = \| x_i - w_{i_1^*} \|_2^2 + \min_p \sum_{k=0}^{P-1} \| w_{p(k+1)} - w_{p(k)} \|_2^2$.

Por fim, o erro combinado (EC) é a média dessa distância sobre as amostras de entrada: $EC = \frac{1}{n}\sum_{i=1}^{n} EC_i$. Segundo [Delgado et al. 2017], à medida que a rede SOM cresce (em termos de número de neurônios m) o erro de quantização diminui, assim como a distância entre os BMUs, sugerindo forte correlação negativa entre esse índice e o tamanho m da rede neural.

2.3. Clusterização dos pesos W do mapa neural

Para clusterização dos pesos da rede neural SOM treinada optamos por avaliar os algoritmos k-médias com a definição do melhor número de agrupamento a partir do método de identificação automática do cotovelo no gráfico da soma da variação total intragrupos para cada valor de c avaliado. Também aplicamos o algoritmo hierárquico aglomerativo, definindo a melhor partição do dendograma a partir do índice Silhouette. Estes dois algritmos de clusterização podem não ser adequados para todos os conjuntos de dados, incluindo aqueles com elevada complexidade em termos de estrutura e distribuição dos dados em cada um dos agrupamentos. Para isto avaliamos três algoritmos baseados na segmentação dos neurônios da rede SOM treinada, considerando a mesma como um grafo não-direcionado. Onde os neurônios i representam os nós do grafo, e as relações de vizinhança que conecta os neurônios os vértices.

Dado uma rede treinada como um grafo $G = (V, E)$, o algoritmo proposto por [Costa and Netto 2003] visa eliminar as arestas consideradas inconsistentes e formar agregados a partir dos nós(neurônios) que permanencerem conectados conforme o Algortimo 1. Neste algoritmo o número de agrupamentos c não é pré-determinado e temos três parâmetros definidos empiricamente.

Algorithm 1 Clusterização proposta por [Costa and Netto 2003]

Require: $G = (V, E)$ – SOM treinada como grafo não-orientado
Require: H_i – Nível de ativação do neurônio i
Require: $d(i, j)$ – Distância entre os pesos w_i e w_j da rede neural
Require: ω – Hiperparâmetro definido empiricamente
for cada par de neurônio adjacente (i, j), a aresta calculada será inconsistente se: **do**

- $d(i, j)$ supera em duas unidades a distância média dos outros neurônios adjacentes a i ou a j.
- i e j tiverem atividade H abaixo de 50% do mínimo permitido H_{min} ou um dos neurônios é inativo $H_i = 0$, dado $H_{min} = \omega H_{med}$, $0.1 \leq \omega \leq 0.6$ e $H_{med} = \frac{n}{m}$.
- A distância entre os centróides dos conjuntos de dados associados aos neurônios i e j excede por duas unidades a aresta $d(i, j)$.

end for
Remove de arestas inconsistentes. Cada aresta inconsistente (i,j) resultará em uma conexão nula no endereço (i, j) da matriz de adjacência.
Atribui rótulos distintos para cada conjunto de neurônios conectados.

O algoritmo proposto por [Silva and Costa 2011] (Algoritmo 2) segue a mesma proposta do particionamento do grafo pela eliminação de arestas consideradas inconsistes proposta por [Costa and Netto 2003]. No entanto, neste caso temos apenas um hiperparâmetro, v, definido empiricamente e que varia no intervalo $0 \leq v \leq 1$.

Algorithm 2 Clusterização do SOM proposto por [Silva and Costa 2011]

Require: $G = (V, E)$ – SOM treinada como grafo não-orientado
$T \leftarrow$ Árvore Geradora Mínima de G usando D como pesos das arestas
for cada aresta (i, j) **do**
Calcule o índice Davis-Bouldin (DBI) para neurônios adjacentes i e j de acordo com a topologia da rede neural
Se $DBI(i, j) \geq v$ então a aresta (i, j) é considerada inconsistente
end for
Associe um rótulo para cada grupo de nós conectados em G

[Silva et al. 2024] propõe a segmentação em k "clusters" baseando-se na distância e na densidade entre neurônios. Primeiro, calcula-se a árvore geradora mínima de G usando D como os pesos das arestas. Em seguida, calcula-se o custo de cada aresta de T, por meio do índice DBI. Por fim, poda-se $k - 1$ arestas de T com os menores custos e atribui-se um rótulo de "cluster" para cada conjuto de nós conectados em T.

O algoritmo proposto por [Silva et al. 2024] primeiro calcula a Árvore de Extensão Mínima (MST), depois define o valor das arestas restantes como o DBI entre os neurônios e poda $k - 1$ arestas com os menores custos, isolando grupos de neurônios. k é um parâmetro a ser definido no início e significa o número de clusters que se deseja obter.

Algorithm 3 Clusterização do SOM proposto por [Silva et al. 2024]

Require: $G = (V, E)$ – SOM treinada como grafo não-orientado
Require: H – Nível de ativação do neurônio
Require: D – Matriz de distância entre os pesos W da rede neural
Require: c – Número desejado de agrupamentos
  $T \leftarrow$ Árvore Geradora Mínima de G usando D como pesos das arestas
  for cada aresta $(u, v) \in T$ **do**
    $cost(u, v) \leftarrow DBI(u, v)$
  end for
  Poda de $c - 1$ arestas em T com os menores custos
  Associe um rótulo para cada grupo de nós conectados em T

3. Bases de dados avaliadas e procedimento experimental

Foram avaliadas quatro bases de dados de benchmark. O primeiro conjunto, Gaussiana, é artificial e simula três conjuntos de dados ($N = 100$) em duas dimensões com distribuição Gaussiana. O segundo conjunto é a base Iris [Fisher 1936] com 150 observações, quatro dimensões, três classes, sendo duas não-linearmente separáveis. A terceira base de dados é a Chainlink [Ultsch et al. 1994], com 1000 observações, tridimensional e com duas classes que representam dois elos de corrente não-lineramente separáveis. A quarta base é a MNIST [Deng 2012], com 10.000 observações, 248 dimensões e 10 classes que representam os dígitos de 0 a 9.

O experimento seguiu as etapas descritas a seguir para cada base de dados analisada:

1. Definição dos hiperparâmetros (raio inicial da função de vizinhança, taxa de aprendizagem, tamanho da grade de neurônios, dimensionalidade e tipo de topologia) que serão avalidos para as redes SOM avaliadas.
2. Logo após, para cada configuração de rede treinada, calcula-se os valores do Erro combinado e da função topográfica. Foram avaliadas as correlações entre os índices de qualidade de preservação topológica (erro combinado e função topográfica) e o tamanho m da rede neural usada na clusterização.
3. Baseado em [Delgado et al. 2017], seleciona-se as redes com menor violação topológica para cada tamanho de grade m, de acordo com o Erro combinado.
4. Em seguida, seleciona-se a rede com a menor violação de topologia, comparando os valores de função topográfica obtidos para diferentes tamanhos de grade geradas pelo passo anterior.
5. Para efeito comparativo também foi selecionada para análise a pior rede SOM conforme o método definido em [Delgado et al. 2017].
6. Aplica-se os cinco algoritmos de clusterização sobre as redes SOM selecionadas.
7. Como todas as bases de dados são rotuladas, para cada clusterização são calculados os índices de qualidade da clusterização NMI, ARI e ACC.
8. Foi aplicado teste estatístico de correlação entre os índices de mensuração da qualidade da preservação topológica (erro cominado e função topográfica) e o tamanho da rede m.

4. Resultados e Discussão

Para todas as bases de dados foram avaliadas redes neurais SOM bi e unidimensional, com grade hexagonal e retangular, função de vizinhança gaussiana, diferentes raios iniciais para essa função $(0.5, 1.0, 1.5, 2.0, 2.5, 3.0)$ e diferentes taxas de aprendizagem para a aprendizagem sequencial $(0.01, 0.05, 0.1, 0.5)$. Para as bases Iris, Gaussian e Chainlink foram variadas as dimensões da rede SOM $((5\times4), (5\times5), (6\times5), (1\times30), (5\times7), (8\times5), (8\times6), (9\times6), (9\times7), (10, 8))$, para a base MNIST foram avaliadas redes um pouco maiores $((8\times8), (15\times10), (20\times20), (25\times30), (30\times30))$.

Para cada uma dessas bases foi escolhida a melhor rede SOM a partir do critério estabelecido por [Delgado et al. 2017]. Para o dataset Gaussiano foi definida como a melhor a seguinte configuração de rede SOM 2D 10×8 com raio inicial igual a 3.0, taxa de aprendizagem igual a 0.5 e topologia retangular. Para as bases Chainlink e Iris foi definida como a melhor rede SOM a configuração bidimensional 10×8 com raio inicial igual a 2.0, taxa de aprendizagem igual a 0.5 e topologia retangular.

A análise do teste de correlação entre os índices de qualidade da rede SOM (Erro combinado e Função topográfica) mostra que a correlação negativa prevista para o Erro combinado foi de menor intensidade e não estatisticamente significativa para todas as bases de dados (Tabela 1). Isto sugere que a componente topológica do indicador (caminho entre o primeiro e segundo BMU) não decresce linearmente à medida que a rede SOM (m) cresce. Enquanto que para o valor para a função topográfica observamos forte correlação positiva e significativa para todas as bases. Isto sugere que uma das componentes ($f_i(k)$ e $f_i(-k)$) desse índice cresce mais que a outra, trazendo como consequência a não anulação do efeito do tamanho da rede SOM.

Tabela 1. Teste de correlação (Cor) entre o Erro combinado (Ec) e a função topográfica (Ft) e o número de neurônios (m). Entre colchetes temos o intervalo de confiança ao nível de 95% para a estatística de correlação.

Dado	Cor(Ec, m)	Cor(Ft, m)
Gaussian	-0.10 [-0.28, 0.08]	-0.83*** [-0.88, -0.77]
Chainlink	-0.46*** [-0.59, -0.31]	-0.79*** [-0.85 -0.71]
Iris	-0.24** [-0.40, -0.06]	-0.80*** [-0.86 -0.73]
MNIST	-0.25 [-0.48, 0.002]	-0.75*** [-0.85 -0.62]

A Tabela 2 mostra os resultados das clusterizações para a melhor e pior redes neurais indicadas pelo método proposto por Delgado et al. (2017) tanto em termos do número c de grupos encontrados como em termos dos índices NMI, ARI e ACC. Os resultados nos permitem afirmar que o método proposto por [Delgado et al. 2017] auxilia na identificação redes neurais SOM que melhor preservam a topologia dos dados de entrada, embora o desempenho do algoritmo de clusterização dependa das características dos dados de entrada. Para a base Gaussian os melhores desempenhos foram obtidos pelos algoritmos de clusterização hierárquico aglomerativo e k-médias, para a base Chainlink o melhor desempenho foi obtido pelo algoritmo proposto por [Silva et al. 2024], para a base Iris o melhor desempenho foi obtido pelo algoritmo k-médias. Para a base MNIST o método hierárquico aglomerativo obteve o melhor resultado.

Tabela 2. Resultados da clusterização dos dados a partir dos cinco algoritmos avaliados usando os índices NMI, ARI e ACC como indicadores da qualidade da partição em c grupos para a melhor e pior redes SOM indicada pelo método proposto por Delgado et al. (2017). Os melhores clusterizadores para cada conjunto de dados estão destacados em negrito.

Base	Algoritmo de clusterização	Melhor rede SOM				Pior rede SOM			
		c	NMI	ARI	ACC	c	NMI	ARI	ACC
Gaussian	**K-means**	**3**	**0.85**	**0.81**	**0.95**	**3**	**0.86**	**0.89**	**0.96**
	H.A.	**3**	**0.87**	**0.83**	**0.96**	3	**0.86**	**0.89**	**0.96**
	Silva et al. (2024)	3	0.42	0.52	0.62	3	0.31	0.17	0.54
	Silva e Costa (2011)	14	0.44	0.57	0.56	4	0.73	0.70	0.82
	Costa e Netto (2003)	2	0.04	0.06	0.44	1	0.00	0.00	0.33
Chainlink	K-means	4	0.25	0.34	0.44	4	0.30	0.20	0.42
	H.A.	9	0.19	0.41	0.28	8	0.50	0.25	0.30
	Silva et al. (2024)	**5**	**0.75**	**0.72**	**0.86**	**2**	**0.21**	**0.09**	**0.65**
	Silva e Costa (2011)	14	0.10	0.31	0.24	4	0.43	0.31	0.60
	Costa e Netto (2003)	4	0.67	0.69	0.73	1	0.00	0.00	0.50
Iris	**K-means**	**3**	**0.76**	**0.77**	**0.91**	**4**	**0.72**	**0.65**	**0.83**
	H.A.	2	0.56	0.73	0.67	2	0.73	0.57	0.67
	Silva et al. (2024)	3	0.56	0.71	0.68	3	0.68	0.54	0.67
	Silva e Costa (2011)	9	0.19	0.33	0.41	4	0.60	0.51	0.70
	Costa e Netto (2003)	8	0.23	0.41	0.54	1	0.0	0.0	0.33
MNIST	K-means	8	0.52	0.68	0.66	5	0.49	0.31	0.45
	H.A.	**10**	**0.60**	**0.75**	**0.72**	**10**	**0.78**	**0.65**	**0.77**
	Silva et al. (2024)	13	0.002	0.09	0.14	3	0.22	0.05	0.20
	Silva e Costa (2011)	60	0.31	0.63	0.35	9	0.60	0.28	0.48
	Costa e Netto (2003)	1	0.00	0.00	0.10	1	0.00	0.00	0.10

5. Conclusões

Conclui-se que o método de seleção da melhor rede SOM proposto por [Delgado et al. 2017] é satisfatório, mas que foi observada correlação negativa entre o tamanho m da rede SOM e a função topográfica $\phi_A^M(0)$, que acaba induzindo como escolha da melhor rede SOM aquela com o maior número de neurônios.

Observou-se que o desempenho do clusterizador dos pesos da rede SOM depende das características do dado de entrada. Sendo que o k-means obteve os melhores resultados para as bases Gaussian e Iris, o método hierárquico aglomerativo para as bases Gaussian e MNIST e o método proposto por [Silva et al. 2024] para a base Chainlink.

Trabalhos futuros incluem a avaliação de outros métodos de avaliação da qualidade do SOM quanto ao ajuste aos dados e à preservação topológica, a aplicação do método de seleção dos hiperparâmetros em bases de dados mais complexas em termos de dimensionalidade e não-linearidade, e avaliação de outros métodos de clusterização dos dados a partir da segmentação da rede neural SOM.

Referências

Appukuttan, A. et al. (2025). Exploring hydrochemical drivers of drinking water quality in a tropical river basin using self-organizing maps and explainable ai. *Water Research*, 284:123884.

Costa, J. A. and Netto, M. L. (2003). Segmentação do SOM baseada em particionamento de grafos. In *VI Brazilian Conference on Neural Networks*, page 451–456.

Delgado, S. et al. (2017). A SOM prototype-based cluster analysis methodology. *Expert Systems with Applications*, 88:14–28.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Hameed, A. A. et al. (2024). Improving the performance of self-organizing map using reweighted zero-attracting method. *Alexandria Engineering Journal*, 106:743–752.

Kaski, S. and Lagus, K. (1996). Comparing self-organizing maps. In *International Conference on Artificial Neural Networks (ICANN)*, ICANN'96, Berlin, Heidelberg. Springer-Verlag.

Kohonen, T. (2001). *Self-Organizing Maps*. Springer Berlin, Heidelberg.

Melo Riveros, N. A., Cardenas Espitia, B. A., and Aparicio Pico, L. E. (2019). Comparison between k-means and Self-Organizing Maps algorithms used for diagnosis spinal column patients. *Informatics in Medicine Unlocked*, 16:100206.

Santos, L. et al. (2021). Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75–88.

Silva, L. and Costa, J. A. (2011). A graph partitioning approach to SOM clustering. In *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, pages 152–159.

Silva, M. A. S. d. et al. (2022). Tracking the connection between Brazilian agricultural diversity and native vegetation change by a machine learning approach. *IEEE Lat. Am. T.*, 20(11):2371–2380.

Silva, M. A. S. d. et al. (2024). A Self-Organizing Map clustering approach to support territorial zoning. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 272–286. Springer Nature Switzerland.

Ultsch, A. et al. (1994). Knowledge extraction from artificial neural networks and applications. *Parallele Datenverarbeitung mit dem Transputer*, pages 148–162.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11(3):586–600.

Villmann, T., Der, R., Herrmann, M., and Martinetz, T. M. (1997). Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 8(2):256–266.