Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 17, Iss. 1, pp. 48-56, 2019

AVALIAÇÃO DA INFLUÊNCIA DE DIFERENTES MEDIDAS DE SIMILARIDADE NA REDE AUTO-ORGANIZÁVEL DE KOHONEN E SUA APLICAÇÃO NA CLASSIFICAÇÃO DE DADOS FÍSICO-QUÍMICOS DA ÁGUA

Carlos A. O. Araujo, Gildson de Jesus, Eduardo Palmeira, Erik da Silva, Fábio Amorim
Universidade Estadual de Santa Cruz
caoaraujo@uesc.br, gildsonj@gmail.com, espalmeira@gmail.com, erikgps@gmail.com, facamorim@gmail.com

Resumo – Sabe-se que cada conjunto de dados apresenta características particulares conforme a sua natureza. Com vista nisso, este artigo avalia a influência de diferentes medidas de similaridades no processo competitivo da rede auto-organizável de Kohonen (SOM, sigla em inglês). Testes numéricos são apresentados para mostrar a eficácia de cada métrica na classificação de três diferentes bases de dados. Além disso, é apresentada uma aplicação da rede SOM, com o uso de diferentes medidas de similaridade, na classificação de dados físico-químicos da água.

Palavras-chave – SOM, Similaridade, Redes Neurais, Métrica, Classificação, Dados Físico-Químicos.

Abstract – It is known that each set of data presents particular characteristics according to their nature. In view of this, this article evaluates the influence of different measures of similarities in the functioning of Kohonen's self-organizing network (SOM). Numerical tests are presented to show the effectiveness of each metric in the clustering of three different databases. In addition, it is apresent an application about of the SOM network, using different measures of similarity, in the classification of physicochemical data of water.

Keywords – SOM, Similarity, Neural Networks, Metric, Classification, Physicochemical Data.

1. INTRODUÇÃO

A rede SOM é um método computacional utilizado na classificação e exploração de dados [1]. Este método tem sido alvo de várias aplicações em diversas áreas nas últimas décadas como: engenharias [2], [3], [4], neurociências [5], [6], ecologia [7], [8], genética [9], [10], geociências [11], química [12] e energia [13]. Isso se deve ao fato de que seu funcionamento é capaz de se ajustar de tal forma que transforma um conjunto de dados de entrada de dimensão elevada em sub-conjuntos de dados de saída uni ou bidimensionais. Todo esse processo é baseado em um aprendizado competitivo, cooperativo e adaptativo, no qual os neurônios têm, na fase inicial, seus pesos sinápticos aleatoriamente escolhidos. Esses neurônios são dispostos em um mapa topológico com a distribuição feita a partir dos conjuntos de vizinhança. Assim, para que os pesos possam ser ajustados com relação aos dados de entrada, concorrem entre si, classificados por alguma medida de similaridade.

Desse modo, no processo competitivo da SOM clássica é utilizada, como medida de similaridade, a medida Euclidiana, cuja métrica é baseada na distância espacial entre dois pontos. Sabe-se, contudo, que alguns problemas de engenharia e de ciências naturais podem estar associados a fenômenos, cuja natureza implique que seus dados se distingam (ou se pareçam), entre si, não por uma distância espacial. Assim, faz-se necessária uma análise mais detalhada de qual medida de similaridade é mais eficaz para uma melhor classificação.

Sendo assim, uma das propostas deste artigo é estudar a influência de medidas de similaridade de naturezas diferentes, a saber: as métricas de Fu [14], de Tanimoto [15], a métrica baseada em Grafo Fuzzy [16], e a métrica Euclideana, no processo competitivo da SOM, com o intuito de investigar se há diferença no uso de diferentes tipos de métricas na rede SOM. Para tanto, foram realizados experimentos numéricos com três diferentes bases de dados: *Glass Identification, Wine, Wine Quality*.

A outra proposta deste artigo é a aplicação da rede neural SOM na classificação temporal de dados físico-químicos da água. No trabalho de Santos Júnior [17] foi realizada uma avaliação das características físico-químicas da água do Rio Grande, localizado na cidade de Barreiras-BA. As amostras foram coletadas em dois pontos distintos do rio e em dois períodos temporais diferentes: período seco e período chuvoso. O objetivo aqui foi utilizar a rede SOM para classificar os dados conforme os períodos temporais em que foram coletados. Na aplicação foram usadas as diferentes métricas no processo competitivo da SOM. Experimentos numéricos foram realizados comparando a eficácia da rede SOM com cada métrica utilizada.

Vale salientar que estudos sobre a influência de diferentes tipos de medida de similaridade no processo competitivo da rede neural SOM, até onde vai o conhecimento dos autores, não foi encontrado na literatura.

2. PRELIMINARES

Nesta seção, serão apresentados a rede neural SOM e as medidas de similaridade que serão utilizadas no processo competitivo da rede SOM, com suas respectivas particularidades.

2.1 REDE AUTO-ORGANIZÁVEL DE KOHONEN (SOM)

Conhecido por ser um método de rede neural artificial (RNA), a rede SOM proposta por Kohonen [**?**] é um diagrama de agrupamento capaz de classificar um conjunto de entrada de dados de dimensão elevada em sub-conjuntos de dados de saída, mantendo as suas características condensadas em um grupo organizado topologicamente de dimensão reduzida. A seguir, é apresentado o algoritmo da rede SOM.

Após a inicialização dos pesos sinápticos, aos quais são atribuídos pequenos valores gerados aleatoriamente, o algoritmo é executado através dos seguintes passos:

1. *Competição* – Para cada padrão de entrada, os neurônios calculam seus respectivos valores de uma função discriminante (medida de similaridade), a qual fornece a base para competição entre os neurônios;
2. *Cooperação* – O neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados, fornecendo assim a base para a cooperação entre os neurônios vizinhos;
3. *Adaptação Sináptica* – Este mecanismo permite que os neurônios excitados aumentem seus valores individuais da função discriminante em relação ao padrão de entrada através de ajustes aplicados a seus pesos sinápticos.

No Passo 1 que ocorre o processo competitivo da rede SOM. É nesta etapa que são usadas as medidas de similaridade com o objetivo de medir o quão próximo estão os dados de entrada e os pesos neurais.

2.2 MEDIDAS DE SIMILARIDADE

As medidas de similaridades são essenciais no processo de classificação, pois são ferramentas responsáveis por quantificar a semelhança entre os dados. Elas são utilizadas no processo competitivo da rede SOM.

Dados $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_n\}$ e $Z = \{z_1, z_2, \ldots, z_n\}$, Goshtasby [18] conceitua a similaridade entre X e Y, $S(X, Y)$ como uma medida que quantifica o grau de correlação entre eles. Assim, S será considerada uma medida de similaridade se satisfaz às seguintes condições [15]:

1. $S(X, X) = S_0$;
2. $-\infty < S(X, Y) \leq S_0 < +\infty$;
3. $S(X, Y) = S_0$, se e somente se $X = Y$;
4. $S(X, Y) = S(Y, X)$;
5. $S(X, Y)S(Y, Z) \leq [S(X, Y) + S(Y, Z)]S(X, Z)$.

Nesse sentido, a seguir, serão apresentadas as medidas de similaridade Euclidiana, de Fu, de Tanimoto e Grafo Fuzzy.

2.2.1 MEDIDA DE SIMILARIDADE EUCLIDIANA

A métrica Euclidiana usual essencialmente mede a distância em linha reta de dois pontos em um espaço de dimensão R^n. Com isso, pode-se definir a similaridade Euclidiana da seguinte forma:

$$S_E(X, Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

Assim, quanto menor for o valor de $S_E(X, Y)$ mais similares serão os pontos, ou seja, mais próximos eles estão um do outro.

2.2.2 MEDIDA DE SIMILARIDADE DE FU

Essa medida de similaridade se mostrou muito útil para algumas aplicações [14], sendo definida como se segue:

$$S_F(X, Y) = 1 - \frac{||X - Y||}{||X|| + ||Y||} \tag{2}$$

Onde $||X - Y|| = S_E(X, Y)$ e $||X|| = \sqrt{\sum_{i=1}^{n} x_i^2}$. Assim definido, o intervalo dessa função varia entre 0 e 1 e a função mede as diferenças qualitativa e quantitativa entre dois vetores [14]. Dessa forma, $S_E(X, Y) = 1$ (valor máximo) quando $X = Y$, tendo o seu valor mínimo (0) se $X = -Y$.

2.2.3 MEDIDA DE SIMILARIDADE DE TANIMOTO

A medida de Tanimoto é definida com se segue [15]:

$$S_T(X,Y) = \frac{X^T Y}{||X||^2 + ||Y||^2 - X^T Y}. \tag{3}$$

Em tal perspectiva, segundo Goshtasby [18], a medida de Tanimoto é proporcional ao produto interno de X e Y e inversamente proporcional às somas das distâncias Euclidianas ao quadrado e do produto interno de X e Y.

Em um emprego mais amplo, a medida de Tanimoto, utilizada para vetores de valores discretos e inspirada na comparação de conjuntos, é apresentada da seguinte forma:

$$\frac{n_{X\cap Y}}{n_X + n_Y - n_{X\cap Y}} = \frac{n_{X\cap Y}}{n_{X\cup Y}}. \tag{4}$$

No caso acima, X e Y são dois conjuntos e n_X, n_Y, $n_{X\cap Y}$ e $n_{X\cup Y}$ são o número de elementos de X, Y, $X \cap Y$ e $X \cup Y$, respectivamente.

2.2.4 MEDIDA DE SIMILARIDADE POR GRAFO FUZZY

A medida de similaridade por grafo fuzzy é uma medida fuzzy que vem acompanhada com a teoria de grafos. Essa métrica está baseada na conexidade fuzzy dos vértices que ligam cada aresta dos grafos. As definições a seguir são fundamentais para a construção da medida por grafos fuzzy [16]:

Definição 1 (Grafo Fuzzy)*Um grafo $\tilde{G} = \{V, A\}$ é considerado fuzzy se V é o conjunto de vértices, A é uma relação fuzzy simétrica e μ_A é uma função que associa cada aresta a um valor no intervalo [0, 1].*

A Figura 1 apresenta um exemplo de grafo fuzzy.

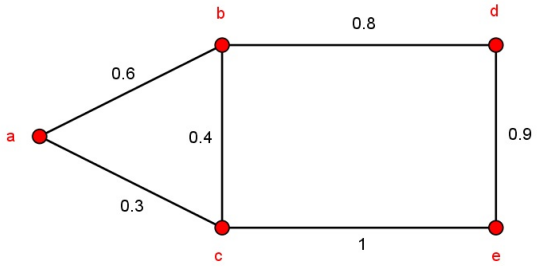

Figura 1: Exemplo de Grafo Fuzzy

Definição 2 (Caminho em um grafo fuzzy) *Um caminho C em um grafo fuzzy $\tilde{G} = \{V, A\}$ é uma sequência de vértices distintos $x_1, x_2, \ldots, x_n$ tal que $\mu(x_i, x_{i+1}) > 0, i = 1, 2, \ldots, n$.*

Definição 3 (Comprimento de um caminho) *Seja C um caminho que passa pelos vértices $x_1, x_2, \ldots, x_n$ em um grafo fuzzy $\tilde{G} = \{V, A\}$.*

Então, o comprimento de C pode ser definido por:

$$\rho(C) = \sum_{i=1}^{n} \frac{1}{\mu(x_i, x_{i+1})}. \tag{5}$$

Definição 4 (Distância fuzzy entre dois vértices) *A distância fuzzy entre dois vértices distintos x_i e x_j, $\delta(x_i, x_j)$, pode ser definida como o menor comprimento de todos os caminhos que ligam esses dois vértices.*

Com base nas definições acima a medida de similaridade por Grafo Fuzzy, entre um certo dado e um neurônio da rede neural SOM, será dada obedecendo os seguintes passos:

1. Formar um grafo entre o dado e todos os neurônios da rede;
2. Calcular o peso de cada uma das arestas (grau de similaridade) que ligam o dado aos neurônios e os neurônios entre si através das seguintes equações:

$$\mu(x_i, w_j) = 1 - S_E(x_i, w_j), \tag{6}$$
$$\mu(w_j, w_k) = 1 - S_E(w_j, w_k), \tag{7}$$

sendo $i = 1, \ldots, t$; $j, k = 1, \ldots, N$ e S_E dado como na equação (1);

3. Calcular o comprimento de todos os possíveis caminhos que ligam o dado ao neurônio segundo a equação (5);

4. Determinar a distância fuzzy entre os dois vértices (dado e neurônio).

3. EXPERIMENTOS NUMÉRICOS

Nesta seção, a rede neural SOM foi implementada para as métricas apresentadas na Seção 2.2, estas foram testadas nas bases de dados *Glass Identification*, *Wine* e *Wine Quality*.

Com o objetivo de comparar as métricas, foram utilizados na SOM os mesmos conjuntos de treinamento e conjuntos de validação em todas as implementações. Foram utilizados também os mesmos vetores de pesos dos neurônios, sendo que estes foram inicializados aleatoriamente. Em todos os experimentos foram utilizados 16 neurônios dispostos em um mapa topológico 4×4. A taxa de aprendizagem utilizada foi $\gamma = 0,01$ para todas as métricas. Os algoritmos foram desenvolvidos utilizando o software MATLAB(R), versão R2015a (8.5.0.197613) 64-bit, instalado no Sistema Operacional Windows 10 Home (64 bits) num computador com processador Intel(R) Core(TM) i5-3317U com 6,00 GB de memória RAM. Logo, a partir dessa configuração, foram usados quatro parâmetros para comparar a eficiência de cada métrica, com relação aos dados que estavam sendo classificados, a saber: a quantidade de épocas para a convergência, a quantidade de neurônios ativados, o tempo de processamento e o percentual de acerto da classificação.

Como pode ser visto nas Tabelas 1, 2 e 3, a SOM utilizando a métrica Grafo Fuzzy apresentou a melhor eficácia de classificação. Em contrapartida, obteve o maior tempo de processamento. Foi constatado que nenhuma métrica conseguiu classificar 100% dos dados do conjunto de validação. As métricas Euclidiana, Fu e Tanimoto apresentaram, praticamente, os mesmos resultados para todos os testes realizados. Observa-se, contudo, que a medida Euclidiana obteve uma menor quantidade de épocas, embora tenha mantido o percentual de acerto menor que a Grafo Fuzzy.

Com isso, a medida de similaridade por Grafo Fuzzy, apesar do maior custo computacional, obteve melhor desempenho na classificação dos dados que todas as outras métricas, nas três bases de dados testadas. Isso se deve, provavelmente, à natureza dos dados e à capacidade da teoria fuzzy em captar as nuances de cada dado.

Tabela 1: Época $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para a base de dados *Glass Identification*

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	4653	80,119	59,09
Fu	4654	131,447	59,09
Grafo Fuzzy	1933	436,013	63,64
Tanimoto	4654	193,522	59,09

Tabela 2: Época $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para a base de dados *Wine*

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	396	7,524	83,33
Fu	403	11,182	83,33
Grafo Fuzzy	315	70,071	86,67
Tanimoto	403	16,473	83,33

Tabela 3: Época $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para a base de dados *Wine Quality*

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	64	9,329	41,67
Fu	68	14,466	41,88
Grafo Fuzzy	61	104,023	47,08
Tanimoto	69	22,197	41,88

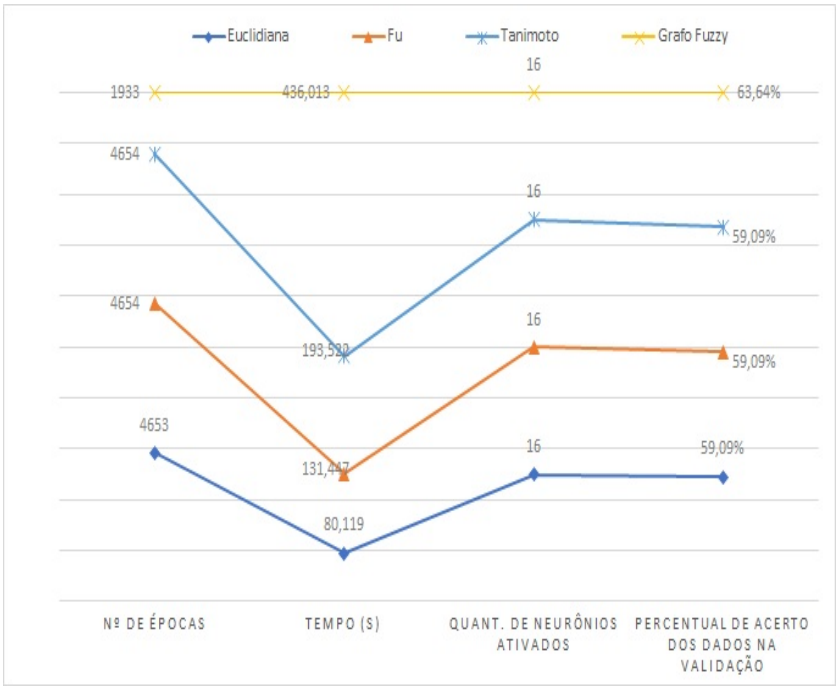

Figura 2: Época × Tempo de processamento × Quantidade de neurônios ativados × Percentual de acerto na classificação para as bases de dados *Glass Identification*

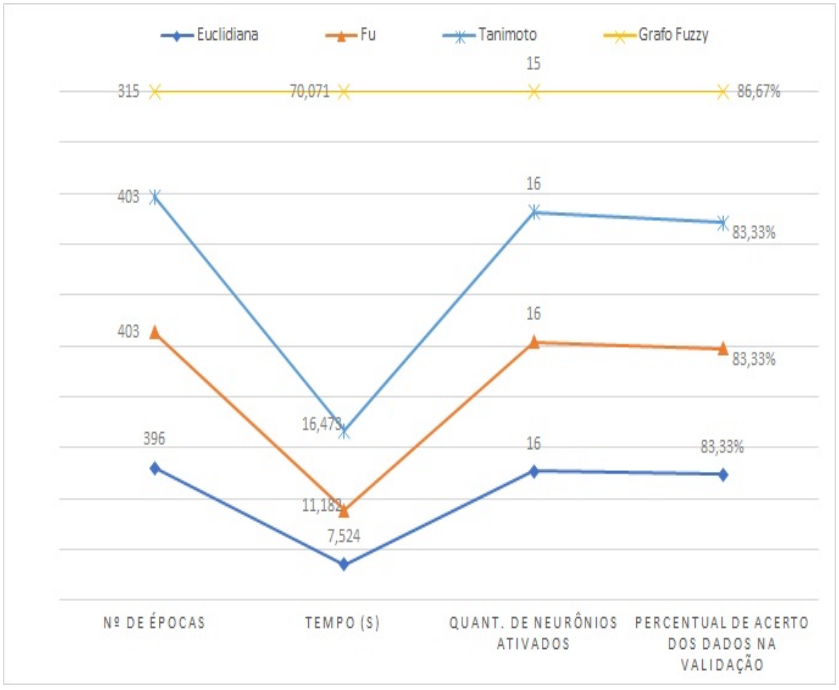

Figura 3: Época × Tempo de processamento × Quantidade de neurônios ativados × Percentual de acerto na classificação para as bases de dados *Wine*

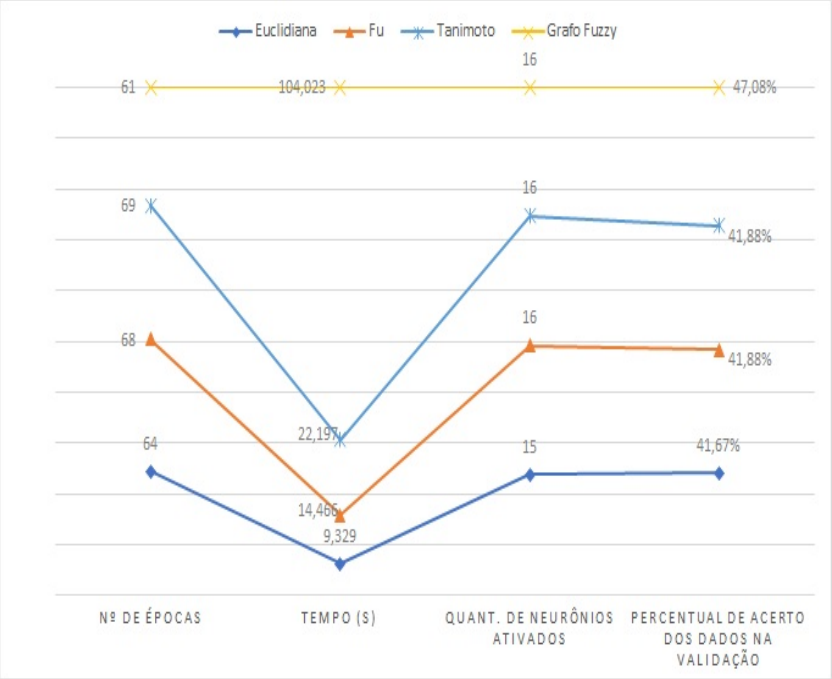

Figura 4: Época × Tempo de processamento × Quantidade de neurônios ativados × Percentual de acerto na classificação para a base de dados *Wine Quality*

4. APLICAÇÃO AOS DADOS FISICO-QUÍMICOS DA ÁGUA

Nesta seção, será realizada uma aplicação da rede SOM, com as diferentes métricas da Seção 2.2, para classificar dados físico-químicos da água, coletados no Rio Grande, localizado na cidade de Barreiras, Bahia. Como pode ser visto em [17], as amostras foram coletadas entre setembro e outubro de 2010 (período seco) e março e abril de 2011 (período chuvoso). E foram considerados dois pontos distintos de coletas no Rio Grande: a estação Ribeirão e a estação Vila Nova. Os parâmetros físico-químicos analisados foram: pH (potencial Hidrogeniônico), temperatura, condutividade elétrica, Oxigênio dissolvido, MPS (material particulado em suspensão), Alumínio, Bário, Cálcio, Ferro, Potássio, Magnésio e Manganês.

Em cada experimento, foram utilizados os mesmos conjuntos de treinamento, conjuntos de validação e vetores de pesos dos neurônios, com o intuito de comparar as eficácias das métricas. Foram utilizados também 16 neurônios dispostos em um mapa topológico 4×4. A taxa de aprendizagem utilizada foi $\gamma = 0,01$ para todas as métricas.

As amostras foram separadas em dois grupos a saber: 48 exemplares do período seco e 48 exemplares do período chuvoso, dos quais 24 correspondem a estação Ribeirão e os outros 24 a estação de Vila Nova.

Como pode ser visto nas Tabelas 4 e 5, os resultados mostram que a métrica de Fu apresentou maior eficiência que as outras. No período chuvoso, a métrica de Fu obteve o maior percentual de acerto na classificação e o menor número de épocas e tempo de processamento, ou seja, maior eficiência com menor custo computacional. No período seco, obteve o maior percentual de acerto na classificação, porém obteve o maior número de épocas. Sendo assim, para esta aplicação, a rede neural SOM, com a métrica de Fu sendo usada no processo competitivo é mais eficiente.

Tabela 4: Época × Tempo de processamento × Percentual de acerto na classificação para os dados físico-químicos da água no período chuvoso

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	2127	10,146	75,00
Fu	828	5,838	83,33
Grafo Fuzzy	954	53,235	83,33
Tanimoto	828	8,198	83,33

Tabela 5: Época × Tempo de processamento × Percentual de acerto na classificação para os dados físico-químicos da água no período seco

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	1012	4,864	91,67
Fu	1013	7,093	91,67
Grafo Fuzzy	1011	55,827	91,67
Tanimoto	1013	10,574	91,67

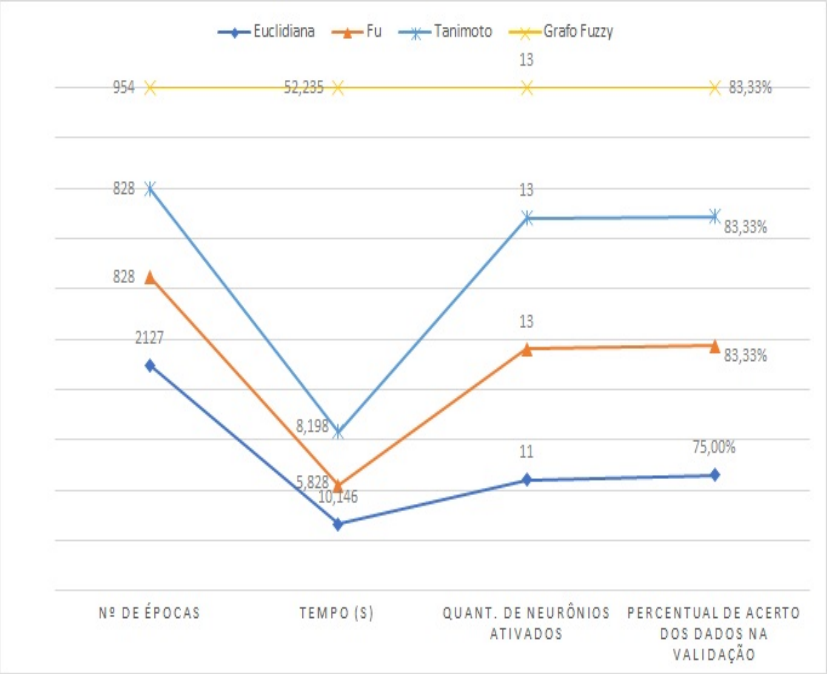

Figura 5: Época × Tempo de processamento × Quantidade de neurônios ativados × Percentual de acerto na classificação para os dados físico-químicos da água no período chuvoso

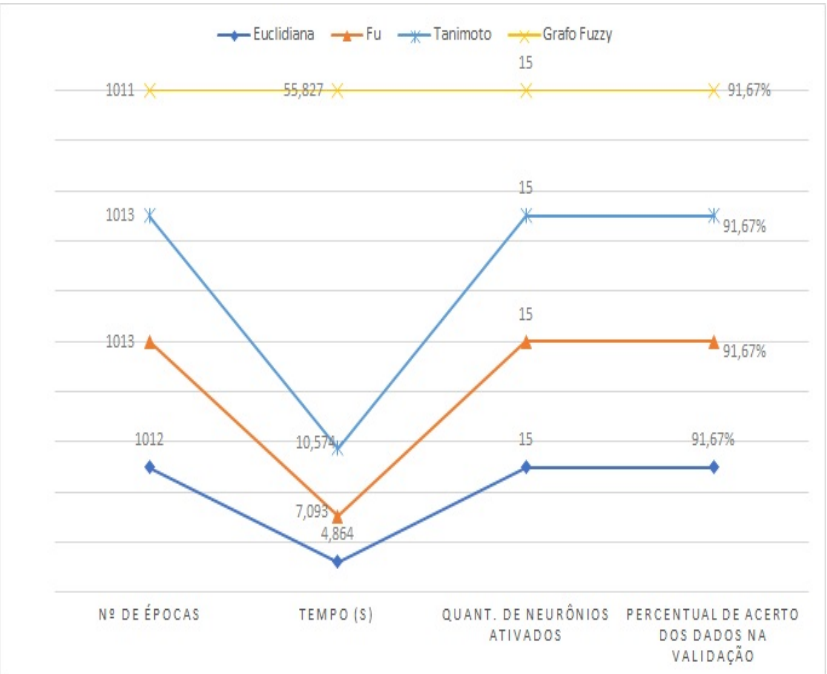

Figura 6: Época × Tempo de processamento × Quantidade de neurônios ativados × Percentual de acerto na classificação para os dados físico-químicos da água no período seco

5. CONSIDERAÇÕES FINAIS

Neste artigo, foi avaliada a influência de diferentes medidas de similaridade na rede neural SOM. Primeiramente, as métricas de Fu, de Tanimoto, grafo fuzzy e Euclideana foram testadas numericamente através de três bases de dados, a saber: *Glass Identification*, *Wine* e *Wine Quality*. Os experimentos numéricos mostraram que a rede SOM com a medida de similaridade por grafos fuzzy obteve maior eficácia para classificar os dados para todas as três bases de dados analisadas e, além disso, apresentou a menor quantidade de épocas, porém com um maior custo computacional. Logo após, foi realizada uma aplicação da rede SOM na classificação temporal de dados físico-químicos da água do Rio Grande. Neste experimento, também a rede SOM foi testada usando os diferentes tipos de medida de similaridade no seu processo competitivo e, foi constatado que, a SOM com a métrica de Fu, para este caso, obteve melhor eficácia na classificação. Portanto, com estes resultados, chega-se à conclusão de que a depender da natureza dos dados, o tipo de medida de similaridade usado no processo competitivo da rede SOM faz diferença na eficiência da classificação dos dados.

REFERÊNCIAS

[1] T. Kohonen. "Self-organized formation of topologically correct feature maps". *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[2] J. S. Torrecilla, J. C. Cancila, G. Matute, P. Díaz-Rodriguez and A. I. Flores. "Self-organizing maps basea no chaotic parametrers to detect adulterations of extra olive oil with inferior edible oils". *Journal of Food Engineering*, vol. 118, pp. 400–405, 2013.

[3] M. A. Valle, G. A. Ruz and V. H. Masías. "Using self-organizing maps to model turnover of sales in a callcenter". *Applied Soft Computing*, vol. 73, pp. 763–774, 2017.

[4] A. N. Chabane, N. Islama and Benoit. "Incremental clustering of sonar images using self-organizing maps combined with fuzzy adaptive resonance theory". *Ocean Engineering*, vol. 142, pp. 133–144, 2017.

[5] P. A. Mei, C. de Carvalho Carneiro, S. J. Fraser, L. L. MIn and F. Reis. "Analysis of neoplastic lesions in magnetic resonance imaging using self-organizing maps". *Journal of the Neorological Science*, vol. 359, pp. 78–83, 2015.

[6] W. L. Chang, L. M. Pang and K. M. Tay. "Application of self-organizing map to failure modes and affects analysis methodology". *Neurocomputing*, vol. 249, pp. 314–320, 2017.

[7] D.-H. Kim, W.-S. Cho and T.-S. Chon. "Self-organizing map and species abundance distruibuition of stream benthic macroinvertebrates in revealing community patterns in different seasons". *Ecological Informatics*, vol. 17, pp. 14–29, 2013.

[8] Y. Nam, S.-H. Koh, S.-J. Jeon, J. Youn, Y.-S. Park and W. I. Choi. "Hazard rating of coastal pine forests for a black pine bas scale using self-organizing map (SOM) and random forest approaches". *Ecological Informatics*, vol. 29, pp. 206–213, 2015.

[9] S. Tapan and D. Wang. "A Further Study on Mining DNA Motifs Using Fuzzy Self-Organizing Maps". *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, pp. 113–124, 2016.

[10] S. S. Ray, A. Ganivada and S. K. Pal. "A Granular Self-Organizing Map for Clustering and Gene Selection in Microarray Data". *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, pp. 1890–1906, 2016.

[11] H. Kalinić, H. Mihanović, S. Cosoli and I. Vilibić. "Sensitivity of Self-Organizing Map surface current patterns to the use of radial vs.Cartesian input vectors measured by high-frequency radars". *Computers and Geosciences*, vol. 84, pp. 29–36, 2015.

[12] C. G. a Novaes, I. L. da Silva Romão, B. G. Santos, J. P. Ribeiro, M. A. Bezerra and E. G. ao Paranhos da Silva. "Screening of Passiflora L. mineral content using principal component analysis and Kohonen self-organizing maps". *Food Chemistry*, vol. 233, pp. 507–513, 2017.

[13] M. Liukkonen and T. Hiltunen. "Adaptative monitoring of emissions in energy boilers using selforganizing maps: An application to a biomass-fired CFB (circulating fluidized bed)". *Energy*, vol. 73, pp. 443–452, 2014.

[14] L. Fu, M. Yang, R. Braulan and N. Benson. "Real-time Adaptive Clustering of Flow Cytometric Data". *Pattern Recognition*, vol. 26, pp. 365–373, 1993.

[15] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier, San Diego, second edition, 2003.

[16] T. S. Silva. "Avaliação da Influência de uma Medida de Similaridade Fuzzy na Clusterização Hierárquica Baseada na Conexidade de Grafos Fuzzy". Master's thesis, Universidade Estadual de Santa Cruz, Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia, Ilhéus, Março 2016.

[17] C. G. P. dos Santos Júnior. "Avaliação Temporal dos Parâmetros Físico-Químicos da Água e da Concentração de Metais em Material Particulado do Rio Grande, Barreiras, Bahia". Master's thesis, Universidade Federal da Bahia, Programa de Pós-Graduação em Química, Salvador, 2014.

[18] A. A. Goshtasby. "Similarity and Dissimilarity Measures". In *Imagem Registration: Principles, Tools and Methods*, edited by A. A. Goshtasby, pp. 7–66. Springer, London, 2012.