



DINCON 2017

CONFERÊNCIA BRASILEIRA DE DINÂMICA, CONTROLE E APLICAÇÕES

30 de outubro a 01 de novembro de 2017 – São José do Rio Preto/SP

Avaliação da Influência de Diferentes Medidas de Similaridade na Rede Auto-Organizável de Kohonen

Carlos A. O. Araujo¹

Programa de Pós-graduação em Modelagem Computacional em Ciências e Tecnologia, UESC, Ilhéus, BA.

Gildson Q. de Jesus²

Programa de Pós-graduação em Modelagem Computacional em Ciências e Tecnologia, UESC, Ilhéus, BA.

Eduardo S. Palmeira³

Programa de Pós-graduação em Modelagem Computacional em Ciências e Tecnologia, UESC, Ilhéus, BA.

Resumo. Sabe-se, que cada conjunto de dados apresenta características particulares conforme a sua natureza. Com vista nisto, este artigo avalia a influência de diferentes medidas de similaridades no funcionamento da rede auto-organizável de Kohonen (SOM). Testes numéricos são apresentados para mostrar a eficácia de cada métrica na clusterização de três diferentes bases de dados.

SOM, Similaridade, Redes Neurais, Métrica

1 Introdução

O SOM é um método computacional utilizado na clusterização e exploração de dados. 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 [6]. 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

¹caoaraujo@uesc.br

²gildsonj@gmail.com

³espalmeira@uesc.br

que os pesos possam ser ajustados com relação aos dados de entrada, concorrem entre si, classificados por alguma medida de similaridade.

Na 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 [5], [4] e de ciências naturais [10] 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 necessário uma análise mais detalhada de qual medida de similaridade é mais eficaz para uma melhor clusterização.

Baseado nisto, neste trabalho, é proposto um estudo sobre a influência das métricas de Fu [2], de Tanimoto [9] e Grafo Fuzzy [9], além da Euclideana, na eficácia do processo de clusterização de dados da SOM. Nos experimentos numéricos são apresentados a eficácia de cada métrica aplicada as bases de dados: Glass Identification, Wine, Wine Quality.

2 Preliminares

2.1 Rede Auto-Organizável de Kohonen (SOM)

Conhecido por ser um método de rede neural artificial (RNA), o SOM proposto por Kohonen [7]. a rede de Kohonen é um diagrama de agrupamento, esta é 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 SOM.

Após a inicialização dos pesos sinápticos, onde são atribuídos a eles 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.

2.2 Medidas de Similaridade

As medidas de similaridades são essenciais no processo de clusterização de dados, pois são ferramentas responsáveis por quantificar a semelhança entre dados. A medida de similaridade é responsável pelo processo competitivo da rede SOM.

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:

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

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 [2], sendo definida como se segue:

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

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 [2]. 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 [3]:

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

Dessa forma, segundo Goshtasby [1], a medida de Tanimoto é proporcional ao produto interno de X e Y e inversamente proporcional às somas das distâncias Euclidianas ao quadrados 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:

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

Onde, 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. Esta 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 [9]:

Definição 2 (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:

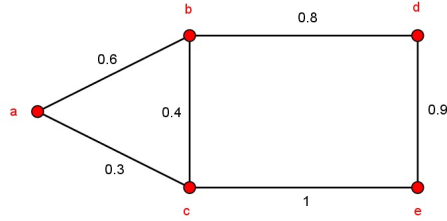


Figure 1: Exemplo de Grafo Fuzzy

Definição 3 (Caminho em um grafo fuzzy) Um caminho C em um grafo fuzzy $\tilde{G} = \{V, A\}$ é uma sequencia de vértices distintos $x_1, x_2, \dots, x_n$ tal que $\mu(x_i, x_{i+1}) > 0, i = 1, 2, \dots, n$.

Definição 4 (Comprimento de um caminho) Seja C um caminho que passa pelos vértices $x_1, x_2, \dots, x_m$ em um grafo fuzzy $\tilde{G} = \{V, A\}$.

Então, o comprimento de C pode ser definido por:

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

Definição 5 (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:

$$\mu(x_i, w_j) = 1 - S_E(x_i, w_j), \quad (6)$$

$$\mu(w_j, w_k) = 1 - S_E(w_j, w_k), \quad (7)$$

sendo $i = 1, \dots, t$; $j, k = 1, \dots, N$ 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

Os testes foram realizados considerando os mesmos dados de entrada no que diz respeito aos vetores de pesos dos neurônios, conjunto de treinamento, conjunto de validação, mapa topológico e conjunto de vizinhança. Com isso, cada métrica apresentada foi testada nas bases de dados *Glass Identification*, *Wine* e *Wine Quality* e alguns dos seus resultados podem ser avaliados a seguir.

Conforme as informações dadas nas Tabelas 1, 2 e 3, temos que a métrica Grafo Fuzzy apresentou a melhor eficácia, mesmo tendo o maior tempo de processamento. Temos também 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, porém a medida Euclidiana obteve uma menor quantidade de épocas, mas manteve o percentual de acerto menor que a Grafo Fuzzy.

Portanto, foi constatado durante os testes, que a medida de similaridade por Grafo Fuzzy, apesar do maior custo computacional, obteve melhor desempenho na clusterização dos dados que as outras medidas nas três bases de dados testadas. Isto se deve talvez, a capacidade da teoria fuzzy de captar as nuances de cada dado.

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	4653	62,976	59,09
Fu	4654	98,197	59,09
Grafo Fuzzy	1933	284,278	63,64
Tanimoto	4654	132,222	59,09

Table 1: Época (iterações) $\times$ 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	396	5,475	83,33
Fu	403	8,526	83,33
Grafo Fuzzy	315	45,345	86,67
Tanimoto	403	12,104	83,33

Table 2: Época (iterações) $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para a base de dados *Wine*

Métrica	Épocas	T. de Proces. (s)	% de Acerto
Euclidiana	64	6,787	41,67
Fu	68	10,840	41,88
Grafo Fuzzy	61	68,144	47,08
Tanimoto	69	15,216	41,88

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

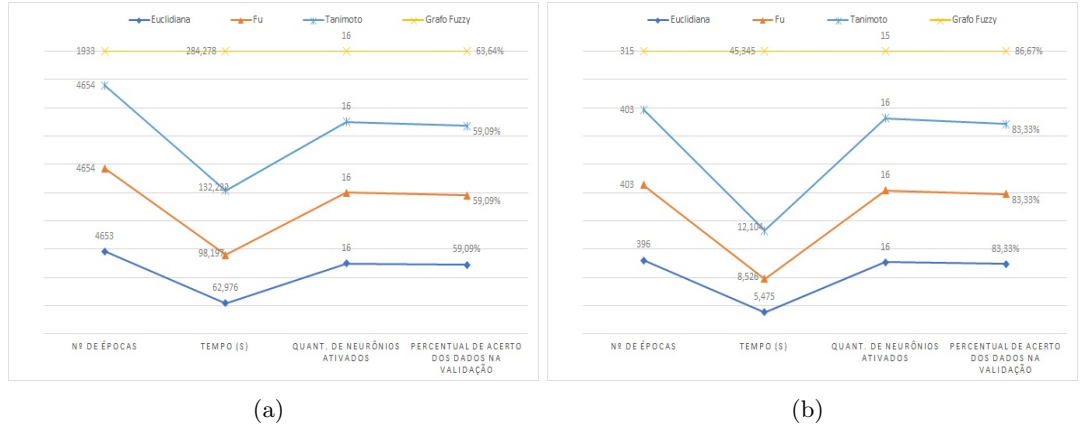


Figure 2: Época (iterações) $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para as bases de dados *Glass Identification* e *Wine*

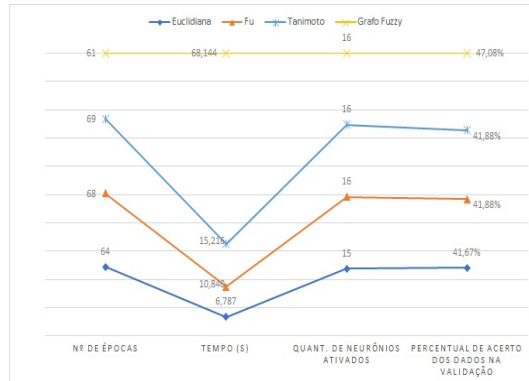


Figure 3: Época (iterações) $\times$ Tempo de processamento $\times$ Percentual de acerto na classificação para a base de dados *Wine Quality*

4 Conclusão

Neste trabalho, avaliou-se a influência de diferentes medidas de similaridade na rede auto-organizável de Kohonen. As medidas foram testadas numericamente através de três

bases de dados, a saber: *Glass Identification*, *Wine* e *Wine Quality*. Os experimentos numéricos mostraram que a medida de similaridade por grafos fuzzy obteve mais eficácia para classificar os dados para todas as bases analisadas e também apresentou a menor quantidade de épocas, porém com um maior custo computacional. Como trabalhos futuros, pretende-se testar outras bases de dados e desenvolver meios para diminuir o custo computacional da medida de similaridade por grafos fuzzy.

Acknowledgment

This work was supported by PPGMC-UESC.

References

- [1] A. A. Goshtasby, *Similarity and Dissimilarity Measures*, Springer, London, (2012).
- [2] L. Fu, M. Yang, R. Braulan and N. Benson, Self-organized formation of topologically correct feature maps, *Pattern Recognition*, vol. 26, 365–373, (1993).
- [3] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Elsevier, San Diego, (2003).
- [4] S. Sankar 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, 1890–1906, (2016).
- [5] 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, 113–124, (2016).
- [6] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, vol. 43, 59–69, (1982).
- [7] T. Kohonen, The self-organizing map, *Neurocomputing*, vol. 21, 1–6, (1998).
- [8] T. Kohonen, Essentials of the Self-Organizing Map, *Neural Networks*, vol. 27, 52–65, (2013).
- [9] T. S. Silva, Avaliação da Influência de uma Medida de Similaridade Fuzzy na Clusterização Hierárquica Baseada na Conexidade de Grafos Fuzzy, *Dissertação de Mestrado em Modelagem Computacional*, UESC, (2016).
- [10] M. Yang, W. Hung and D. Chen, Self-organizing map for symbolic data, *Fuzzy Sets and Systems*, vol. 203, 49–73, (2012).