Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, Vol. 9, Iss. 1, pp. 20-31, 2011
© Sociedade Brasileira de Redes Neurais (SBRN)

UMA NOVA ABORDAGEM PARA VISUALIZAÇÃO E DETECÇÃO DE AGRUPAMENTOS EM MAPAS DE KOHONEN BASEADO EM GRADIENTES DAS COMPONENTES

José Alfredo F. Costa

Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN
E-mail: alfredo@ufrnet.br

Resumo - Visualização e agrupamentos estão entre principais tarefas de mineração de dados. Mapas neurais, como o self-organizing maps (SOM), têm sido bastante utilizados em várias aplicações. Entretanto, etapas de pós-processamento ao treinamento são necessárias para visualização e extração do conhecimento obtido a partir da configuração dos neurônios. Este artigo apresenta uma nova forma de visualização da rede SOM, a GC-matrix, baseada no gradiente das componentes. São apresentadas comparações com a forma tradicional de visualização do SOM, a U-matrix, que é derivada das distâncias entre neurônios. Resultados são apresentados tanto no aspecto de visualização quanto em agrupamentos de dados, para mapas de tamanhos diferentes e vizinhanças finais diferentes. Mostra-se que a GC-matrix apresenta maior estabilidade para escolha de filtragem morfológica, etapa anterior a segmentação. Resultados da aplicação da watershed para a U-matrix e GC-matrix são mostradas, para conjuntos de Gaussianas bivariadas com certo grau de sobreposição.

Palavras-chave- Mapas auto-organizáveis, visualização, agrupamento de dados, redes neurais, reconhecimento de padrões.

Abstract - Visualization and clustering are among the main tasks of data mining. Neural maps, such as self-organizing maps (SOM), have been widely used in a diversity of applications. However, post-processing steps of the training are required for visualization and extraction of knowledge obtained from a trained map. This paper presents a new approach for visualization of neuron activity, the GC-matrix, which is based on the gradient of the components. Comparisons are presented with the U-matrix, which is derived from the distances between neurons. Results of applying the watershed for the U-matrix and GC-matrix are shown for different datasets. Both aspects of data clustering and visualization for maps with different sizes are presented. In our simulations, it is shown also that the GC-matrix is more stable than U-matrix when choosing morphological filtering and segmentation parameters.

Keywords- Self-organizing maps, visualization, data clustering, neural networks, pattern recognition.

1. Introdução

Os avanços tecnológicos nos sistemas de aquisição e armazenamento de dados, aliado a queda dos custos dos dispositivos, estão oferecendo grandes oportunidades para o desenvolvimento e aplicação de novos métodos de reconhecimento de padrões e mineração de dados. Exemplos incluem sistemas nas mais variadas áreas de aplicações, como por exemplo, bases de dados distribuídas na internet, organização e recuperação de imagens baseada em conteúdo, redes de sensores, análise de dados geoespaciais, bioinformática, entre tantos outros que demandam novos métodos de análise e processamento. Analisar estes dados é uma tarefa difícil não apenas devido ao tamanho das bases de dados, sua complexidade (e dimensionalidade), mas também por problemas de escalonamento e descoberta dos padrões escondidos nas massas de dados.

Técnicas de Mineração de Dados (Witten e Frank, 2005), resultado da combinação de métodos de várias áreas, como estatística, inteligência artificial, agrupamentos, e várias outras, objetivam a identificação de padrões, tendências e anomalias em grandes bases de dados. Dentre as principais técnicas de análise de dados, a análise de agrupamentos objetiva a descoberta dos grupos naturais baseados na similaridade dos dados (Xu e Wunsch, 2009). Entretanto a maioria dos métodos de agrupamento de dados possuem geometrias pré-estabelecidas que impõem uma estrutura aos dados, além de serem susceptíveis a inicializações e escolhas de parâmetros.

Mapas neurais constituem-se em um dos modelos mais importantes na área das redes neurais, combinado aspectos de quantização vetorial com a propriedade de continuidade de funções. Mapas auto-organizáveis (SOM - self-organizing map) têm sido utilizados em várias aplicações, incluindo áreas como reconhecimento de padrões, processamento de sinais, compressão de dados e mineração de dados. Baseado em aprendizado competitivo e não-supervisionado (Kohonen, 2001), o SOM possui propriedades importantes como a capacidade de aproximar o espaço de entrada, ordenação topológica e

© Sociedade Brasileira de Redes Neurais (SBRN)

casamento de densidade, aliadas a simplicidade do modelo e a facilidade de implementação do seu algoritmo. O SOM tem sido amplamente utilizado para problemas de visualização e classificação não-supervisionada (Kohonen, 2001), porém vários desenvolvimentos ainda são necessários para obtenção de soluções de problemas complexos.

O método mais comum de visualização do SOM é através da *U-matrix*, que é uma imagem decorrente das distâncias entre neurônios adjacentes (Ultsch, 1993). Este artigo apresenta uma nova forma de visualização do mapa treinado, baseado nos gradientes dos componentes do SOM. Comparações são realizadas entre o novo método e a *U-matrix*. O artigo também ilustra resultados para detecção automática usando SOM e operadores morfológicos para as duas formas de visualização. Resultados usando bases de dados mostram que os resultados em acurácia são similares, porém o método proposto é mais robusto para a escolha de parâmetros na fase de filtragem do que o método convencional. Resultados demonstram a capacidade de transformação de dados multivariados em imagens e sua segmentação, permitindo entendimento da estrutura dos mesmos.

O resultado da aplicação destes algoritmos são regiões conectadas de neurônios que definem no espaço de atributos (entrada) geometrias complexas e não paramétricas. Nos três casos o número de segmentos (ou agrupamentos) é estimado por regras heurísticas e baseada nos dados. Mapas segmentados possibilitam também a classificação de novas amostras. O processo não utiliza informações das classes em nenhum momento do treinamento ou análise (definições dos agrupamentos).

O restante do artigo é organizado da seguinte forma: Seção 2 descreve brevemente SOM e a *U-matrix*. A Seção 3 descreve a proposta da *GC-matrix* enquanto que a seção 4 aborda filtragem e agrupamento através de segmentação morfológica. A Seção 5 apresenta resultados e a Seção 6 conclui o artigo apontando também algumas possibilidades de continuidade deste trabalho.

2. O Mapa de Kohonen e a U-matrix

O SOM define um mapeamento de um espaço *p*-dimensional contínuo para um conjunto finito de vetores referência, ou neurônios, dispostos na forma de um arranjo espacial regular, normalmente bidimensional. O objetivo do treinamento é reduzir a dimensionalidade dos sinais ao mesmo tempo em que se busca preservar ao máximo a topologia do espaço de entrada (Kohonen, 2001).

Cada neurônio i é representado por um vetor de pesos $\mathbf{m}_i = [m_{i1}, m_{i2}, ..., m_{ip}]^T$ onde p é a dimensão dos vetores de entrada. Para cada padrão de entrada um neurônio é escolhido o vencedor, c, usando o critério de maior similaridade,

$$\|\mathbf{x} - \mathbf{m}_c\| = \min_i \{\|\mathbf{x} - \mathbf{m}_i\|\} \tag{1}$$

onde ||.|| representa a distância Euclidiana. Os pesos do neurônio vencedor, bem como os pesos dos neurônios compreendidos em sua vizinhança N_c, são atualizados de acordo com a Equação

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)] \tag{2}$$

onde t indica a iteração, $\boldsymbol{x}(t)$ é o padrão de entrada fornecido de forma aleatória na iteração e $h_{ci}(t)$ é o núcleo de vizinhança em torno do neurônio vencedor. Uma variação do SOM convencional é o algoritmo em lote, que o torna insensível a seqüência de apresentação dos dados em cada época. As contribuições de cada padrão são acumuladas e ao final de cada época que é feita a atualização dos pesos (Kohonen, 2001).

2.1 A U-matrix

A densidade dos neurônios em um mapa treinado é uma aproximação da densidade dos dados (Kaski et al., 2000). Assim, é possível obter informações dos agrupamentos analisando as relações geométricas dos neurônios após o treinamento.

A idéia básica da *U-matrix* é usar a mesma métrica utilizada durante o treinamento para calcular distâncias entre pesos dos neurônios adjacentes (Ultsch, 1993). O resultado é uma imagem $f(x, y)$, na qual as coordenadas de cada pixel (x, y) são derivadas das coordenadas dos neurônios no *grid* do mapa. A intensidade de cada *pixel* na imagem $f(x, y)$ corresponde a uma distância calculada. Pode-se pensar uma imagem como uma superfície em 3D cuja topografia revela a configuração dos neurônios obtida pelo treinamento. Pode-se abstrair vales e montanhas, os primeiros correspondendo a regiões de neurônios similares, enquanto que montanhas refletem a dissimilaridade entre neurônios vizinhos e podem ser associadas a regiões de fronteiras de agrupamentos.

© Sociedade Brasileira de Redes Neurais (SBRN)

Considere um mapa retangular de tamanho $X \times Y$. Seja $[b_{x,y}]$ a matriz de neurônios e $[wi_{x,y}]$ a matriz de pesos. Para cada neurônio em b existem três distâncias d_x, d_y e d_{xy}, a seus vizinhos. Estas distâncias, calculadas no espaço dos pesos, são plotadas em uma matriz U de tamanho $(2X\text{-}1) \times (2Y\text{-}1)$. A *U-Matrix* (Equação 3) é preenchida de acordo com a Tabela 1, onde as abreviações ´*I*´ e ´*P*´ referem-se ao índice ou posição do neurônio, sendo ímpar e par, respectivamente.

$$\begin{bmatrix} du(0,0) & dx(0,0) & du(1,0) & \dots & du(X\text{-}1,0) \\ dy(0,0) & dxy(0,0) & dy(1,0) & \dots & dy(X\text{-}1,0) \\ du(0,1) & dx(0,1) & du(1,1) & \dots & du(X\text{-}1,1) \\ dy(0,1) & dxy(0,1) & dy(1,1) & \dots & dy(X\text{-}1,1) \\ \dots & \dots & \dots & \dots & \dots \\ du(0,Y\text{-}1) & du(0,Y\text{-}1) & du(1,Y\text{-}1) & \dots & du(X\text{-}1,Y\text{-}1) \end{bmatrix} \tag{3}$$

Tabela 1 - Esquema para preenchimento da U-Matrix

i	j	(i,j)	$U_{i,j}$
I	P	$(2x+1, 2y)$	$dx(x,y)$
P	I	$(2x, 2y+1)$	$dy(x,y)$
I	I	$(2x+1, 2y+1)$	$dxy(x,y)$
P	P	$(2x, 2y)$	$du(x,y)$

O cálculo de $du(x, y)$ pode ser feito usando o valor médio ou a mediana dos elementos circunvizinhos.

Vários métodos têm sido propostos para visualização de relações de dados usando SOM, incluindo visualizações de múltiplas componentes, projeções e gráficos 2D e 3D de superfície de matrizes de distância (Kohonen, 2001; Ultsch, 1993; Vesanto, 2000).

3. Visualização do SOM baseado nos gradientes dos componentes

A nova matriz de visualização pode ser obtida a partir dos gradientes das componentes do SOM, denominada *GC-matrix*. Dada a dimensão dos dados, D, que é o mesmo número de elementos do mapa, o objetivo é a obtenção do gradiente de D componentes nas direções do espaço de saída do mapa (dois, na maioria dos casos práticos).

Para um neurônio F com coordenadas (i, j), o valor associado a intensidade é a soma da raiz quadrada do gradiente de componentes. No caso da rede com espaço de saída 2-D (X e Y), podemos expressar como a eq. 4:

$$GC(i,j) = \left\{ \sum_{k=1}^{D} \left[\left(\frac{\partial F_k(i,j)}{\partial x} \right)^2 + \left(\frac{\partial F_k(i,j)}{\partial y} \right)^2 \right] \right\}^{\frac{1}{2}} \tag{4}$$

Para uma melhor visualização e análise de imagem, pode-se usar interpolação bilinear na imagem, de forma que possamos comparar, em tamanho, com a *U-matrix*. Para dimensões superiores do espaço de saída, novos termos na eq. 4 podem ser adicionados. Por exemplo, para a grade de saída 3-D, teríamos $GC(i, j, l)$ com um acréscimo do termo com l no somatório da eq. 4. Equações para espaços de saída do SOM maiores que 3 podem ser também obtidos, de uma forma similar ao *U-array* (Costa eNetto, 2007).

4. Filtragem e agrupamento através de segmentação morfológica

A estratégia de análise de agrupamentos proposta neste trabalho é encontrar uma partição para um conjunto de dados a partir da análise da *U-matrix* [8] e da *GC-matrix,* seguindo os passos do método apresentado por Costa e Netto (1999a, 2001), que emprega o algoritmo de segmentação de imagens, *watershed*, utilizando uma imagem de marcadores para regularizar o processo de segmentação. O resultado da segmentação da *U-matrix* ou da *GC-matrix* produz é associada a regiões de neurônios do SOM, refletindo os diferentes agrupamentos dos dados.

Seja a *U-matrix* (ou *GC-matrix*) de um SOM treinado dada pela imagem *f*, de tamanho *2N-1* × *2M-1*, onde $N \times M$ é o tamanho do mapa. Um passo fundamental é a escolha dos marcadores para a *watershed*. Considere que $[f_{min}, f_{max}] = [0, 255]$ e n_i é 1, ou seja, há 256 níveis de cinza na imagem *f*. Os seguintes passos são efetuados:

1. Filtragem: a imagem f_1 é gerada a partir da abertura e fechamento morfológico por área (Dougherty e Lotufo, 2003), parâmetro *t*, na imagem *f*.
2. Para $k = 1, \ldots, f_{max}$, onde f_{max} é o nível de cinza máximo na imagem f_1, crie as imagens binárias f_2^k correspondendo a conversões de f_1 usando *k* como valor de limiar.
3. Calcule o número de regiões conectadas de f_2^k, para cada valor de *k*, N_{rc}^k.
4. Busca no vetor (ou gráfico) $k \times N_{rc}^k$ a maior seqüência contígua e constante de número de regiões conectadas N_{rc}^k, denotado por S_{max}.
5. A imagem de marcadores será a imagem f_2^j, onde *j* é o valor inicial da seqüência S_{max}.

O passo 1 objetiva melhorar a *U-matrix*, através de filtragem da imagem, removendo poros (área por exemplo, menor ou igual a t pixels). A conectividade é dada pelo elemento estruturante B_c. A abertura por área é definida como

$$area - \gamma_{B_c,t} = \bigcup_{B \in \Xi_{B_c,t}} \gamma_{B_c,t}(f) \tag{5}$$

Onde $\Xi_{B_c,t} = \{X \subset E, X \text{ é } B_c - \text{conectado}, \text{Area}(X) \geq t\}$ e $\gamma_B(f) = \delta_B[\varepsilon_B(f)]$, erosão seguida de dilatação. Eq. 5, remove todos os grãos (isto é, um componentes conectados) com área inferior a um limiar na imagem *f* binária. A conectividade é dada pelo elemento estruturante B_c. Esta função pode ser generalizada para imagens em escala de cinza, aplicando o limiar, sucessivamente, sobre fatias de *f* obtidas a partir dos níveis mais elevados até os mais baixos [10]. De forma semelhante, o fechamento por área (eq. 6), remove todos os poros (ou seja, componentes conectados do *background*) com área inferior a um limiar de uma imagem de *f*. A generalização para imagens em escala de cinza é realizada aplicando o operador binário sucessivamente nas fatias de *f* obtidas a partir dos níveis de limiar superiores aos níveis inferiores (Dougherty e Lotufo, 2003):

$$area - \phi_{B_c,t} = \bigcap_{B \in \Xi_{B_c,t}} \phi_B(f) \tag{6}$$

onde $\phi_B(f) = \varepsilon_B[\delta_B(f)]$, dilatação seguida de erosão.

A filtragem suaviza a imagem e atenua as estruturas menores. Normalmente, o valor de *t* é pequeno, 3 ou 4, para os mapas pequenos, ex. 6x6. Para mapas grandes, ex. 40 x 40, percebe-se que o uso de *t* como metade de uma das dimensões do mapa (*N/2*) têm produzido bons resultados. A escolha influencia resultados de agrupamento do mapa, pois, a filtragem antecede a segmentação. Na maioria dos casos notáveis, os resultados são estáveis, com *t* dentro da faixa 3 a *N/2*. Se *t* é muito elevado ocorrerá redução do número de clusters descobertos. Em princípio, isso não seria um problema para a abordagem completa que trabalha com método hierárquico (Costa e Netto, 2001). Portanto, clusters conectados em um nível podem ser separados em um próximo nível da árvore, com o algoritmo TS-SL-SOM (Costa e Netto, 2001).

A faixa de níveis de cinza neste trabalho foi [0, 255]. Considerando o SOM como um grafo de similaridade, durante a etapa 2 o aumento do nível de cinza de 0 a 255 no limiar da *U-matrix* está relacionada ao aumento do raio de uma zona de influência de cada neurônio no espaço de entrada. Portanto, a partir de zero para a maior distância entre os neurônios vizinhos, 256 níveis são testados na busca de configurações estáveis para fusão das regiões dos neurônios.

O número de regiões conectadas é armazenado no vetor N_{cr}. O gráfico de N_{cr}^k versus *k* ilustra como o número de regiões conectadas (possíveis marcadores para os *clusters*) mudam com o nível de cinza, ou aumento da distância entre neurônios (*k*). O algoritmo procura platôs significativos de N_{cr} na faixa útil da escala de cinza. O algoritmo pode ser considerado como um método baseado em escala. Uma vez definido o platô, o valor inicial k da sequência é escolhido como limiar para gerar os marcadores, embora quaisquer outros valores no platô definido possam ser usados. A idéia da escolha do valor inicial do platô seria ter os menores marcadores possíveis, modificando o mínimo a homotopia da imagem, dando maior flexibilidade para a o algoritmo *watershed* encontrar a melhor partição durante a execução.

Uma vez tendo os marcadores, a segmentação da *U-matrix* (ou da *GC-matrix*) é a aplicação da *watershed* sobre a *U-matrix*, com a mudança da homotopia através do uso dos marcadores e posterior rotulagem das regiões conectadas. O último passo é a cópia dos rótulos da *U-matrix* (ou da *GC-matrix*) para os rótulos dos neurônios associados.

© Sociedade Brasileira de Redes Neurais (SBRN)

5. Resultados

A plataforma de implementação utilizada foi um computador pessoal com processador Intel Core 2 Duo com memória 2 Gb. Os algoritmos foram implementados em Matlab e também foram utilizadas algumas rotinas do SOM Toolbox (SOM Toolbox Team, 2002).

5.1 Mistura de Gaussianas bivariadas

Um conjunto de dados para testes com redes neurais Gaussianas foi proposto em (Hamad et al., 1996). Cinco classes foram geradas contendo cada uma 300 amostras. As cinco populações foram geradas a partir dos vetores de médias (0,0), (1,1), (1, -1), (-1, -1), (-1, 1). A matriz de covariâncias da primeira classe é diagonal, $\Sigma_1 = diag(0.2, 0.2)$ e as outras matrizes de covariâncias, também diagonais, foram obtidas usando $\Sigma = diag(0.05, 0.3)$ rotacionadas com ângulo $\pm \pi/4$. Os dados gerados são apresentados na Figura 1. Nota-se que há sobreposição nas classes, acarretando maior dificuldade de separação dos agrupamentos.

Foram feitas várias simulações utilizando o algoritmo *expectation-maximization* (*EM*) (Xu e Wunsch, 2009). Após determinar o número de classes, simulando valores na faixa de $k = 2$ a 8, usando índices de validação (BIC) (Xu e Wunsch, 2009), o valor mais indicado pelo BIC foi $k = 5$. A acurácia do EM foi 96.53%. No caso do SOM foi utilizado um mapa com dimensão 2 e tamanho do *grid* 12×12, inicializado de forma linear. Após 500 épocas de treinamento obtemos a configuração de neurônios como mostrada na Figura 2. A função de vizinhança utilizada foi Gaussiana, onde o raio inicial usado foi 9, decrescendo até 1, de forma linear. Note que houve uma concentração de neurônios nas regiões de maior densidade de pontos, e que os dados foram escalonados ao intervalo [0, 1].

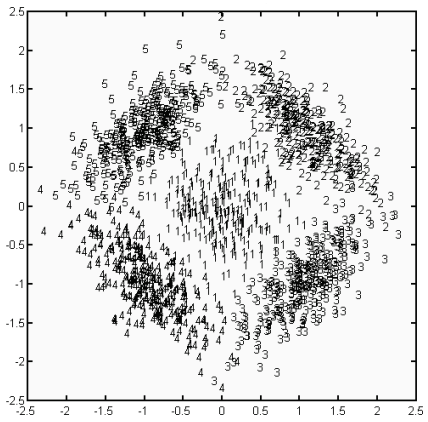

Figura 1: Ilustração do conjunto de dados.

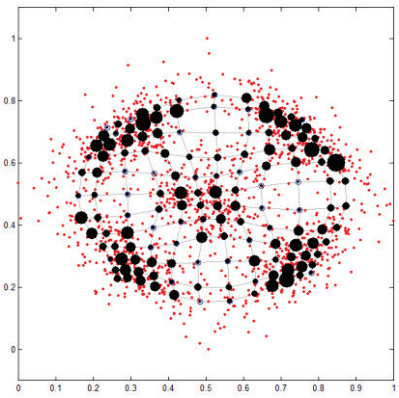

Figura 2: Grid de um SOM com dimensões 12×12 após 500 épocas de treinamento, algoritmo em lote, vizinhança final 1.

As Figuras 3 (a e b) ilustram a *U-matrix* e a *GC-matrix* para um SOM com tamanho 12 × 12, enquanto que as Figuras 4 e 5 ilustram para o caso com SOM com tamanho 18 × 18 (similar ao caso apresentado na Figura 2, apenas com maior número de neurônios para facilitar a visualização). Apesar de detectarmos visualmente a existência de 5 agrupamentos, testes com vários

métodos convencionais de segmentação não apresentaram bons resultados. As Figuras 4 e 5 ilustram a *U-matrix* e a *GC-matrix* para mapas com tamanho 18 × 18 e 25 × 25, respectivamente, em escala de cinza.

Nos dois casos, *U-matrix* e *GC-matrix* (Figuras 3 a 5), foi possível detectar o número correto de agrupamentos, e há pequenas diferenças em relações aos segmentos encontrados, notadamente a parte inferior do cluster central, que no caso da *GC-matrix* se estende mais para baixo. A Tabela 2 ilustra valores de acurácia para diferentes tamanhos de mapas (6x6 a 25x25) e também para vizinhança final no treinamento do SOM em 0 e 1. O parâmetro *t*, relacionado a filtragem, passo 1 do algoritmo apresentado na Seção anterior, foi utilizado o valor inteiro após arredondamento de uma das dimensões do mapa.

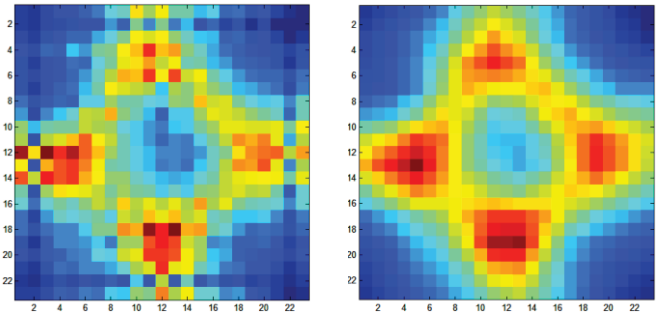

Figura 3: Imagens em pseudo cor. (a) esquerda, *U-matrix*; e (b) direita, *GC-matrix*, obtidas para o SOM com tamanho 12x12.

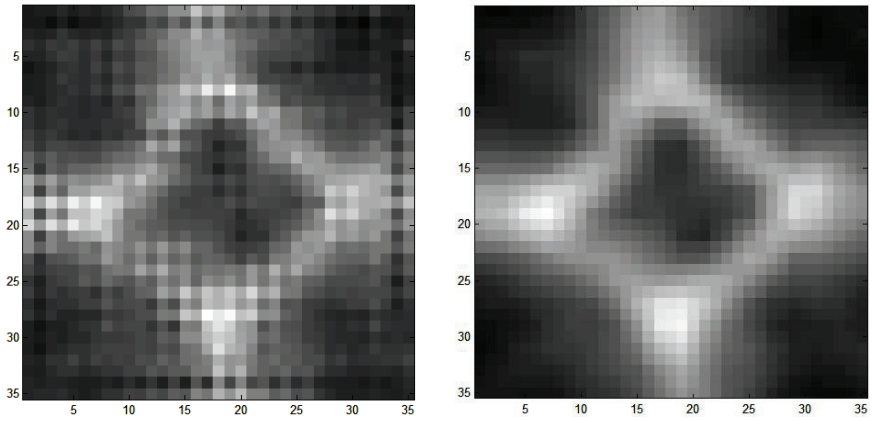

Figura 4: (a) esquerda, *U-matrix*; e (b) direita, *GC-matrix*, obtidas para o SOM com tamanho 18x18.

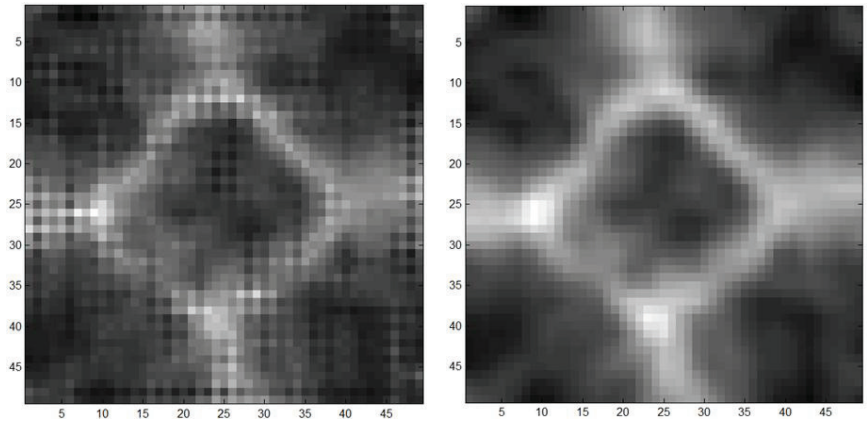

Figura 5: (a) esquerda, *U-matrix*; e (b) direita, *GC-matrix*, obtidas para o SOM com tamanho 25x25.

A Figura 6 ilustra, de forma de superfície, as Figuras 5a e 5b. Nota-se que a *GC-matrix* apresenta maior suavidade, propiciando, visualmente, uma melhor identificação dos cinco agrupamentos. Também, em termos de análise automática, a presença de menos oscilações nos valores da imagem auxilia uma melhor segmentação. Efetuando a análise como descrita na Seção anterior, os resultados obtidos, regiões conectadas de neurônios, são apresentados nas Figuras 7 e 8.

© Sociedade Brasileira de Redes Neurais (SBRN)

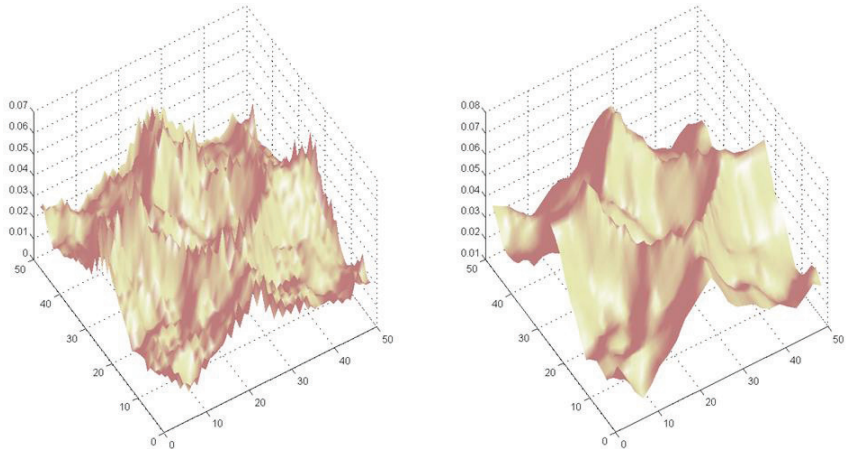

Figura 6: (a) esquerda, *U-matrix*; e (b) direita, *GC-matrix*, obtidas para o SOM com tamanho 25x25.

TABELA 2: ACURÁCIA PARA DIFERENTES TAMANHOS DE MAPAS VIZINHANÇA FINAL DO SOM EM 0 E 1.

	Vizinhança final 1		Vizinhança final 0	
Tamanho do mapa	U	G	U	G
6x6	(a)	(a)	94.7%	94.7%
12x12	94.6%	94.0%	94.9%	94.9%
18x18	95.0%	95.2%	95% (b)	95.2%
25x25	95.7%	95.7%	94.3%	94.8%

Pelos resultados apresentados na Tabela 2, a acurácia obtida a partir do algoritmo usando a *U-matrix* (U) e a *GC-matrix* (G) são equivalentes. No caso do mapa de tamanho 6x6, quando a vizinhança final foi 1, o algoritmo, para os casos U e G não encontraram o número correto de agrupamentos, ficando em 4 clusters, razão pela qual não foram calculados os valores. No caso do mapa 18x18, com vizinhança final 0, o número de clusters encontrado para a maioria dos valores estáveis de t foi diferente de 5. Porém, no intervalo [6, 10], faixa que onde se situa o parâmetro segundo a regra empírica, o número correto de clusters foi encontrado e a acurácia de 95,0% foi obtida.

As Figuras 9 e 10 ilustram, respectivamente para o caso com uso da *U-matrix* e da *GC-matrix*, o número de agrupamentos, ou seja, regiões conectadas, N_{rc}^{k} (no eixo z) versus k, limiar de distância (em x), e t, parâmetro de área do filtro morfológico (no eixo y), para o mapa com tamanho 18 × 18, vizinhança final de treinamento 1.

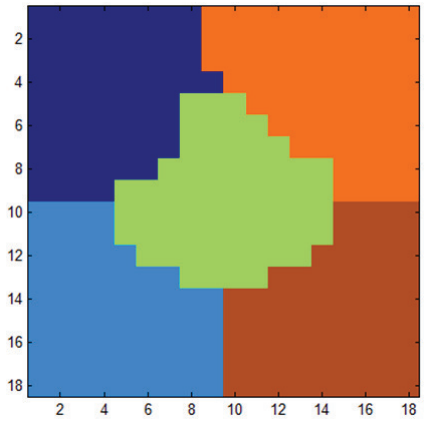

Figura 7: regiões conectadas de neurônios com uso da *U-matrix*.

Nota-se, para o caso da *GC-matrix* (Figura 10), maior tamanho do platô de estabilidade no número correto de agrupamentos (cinco), permitindo maior flexibilidade na escolha dos parâmetros. Também, comparando os resultados da Tabela 2 com os obtidos pelo método de misturas de densidades de probabilidades, algoritmo EM, vemos que diferença é pequena. O resultado é bom, principalmente porque não tivemos que supor que os dados eram provenientes de distribuições Gaussianas, não foram usados métodos para determinar o número de densidades componentes (critérios de informação, baseados em estimativas da

verosimilhança), nem houve necessidade de estimar os vários parâmetros das várias densidades componentes. O método das misturas foi melhor devido à própria estrutura dos dados, que se encaixam perfeitamente no modelo geométrico dos protótipos obtidos. Porém, não há no EM, como ocorre no SOM, a visualização dos agrupamentos em forma de mapa, bastante importante em análises de dados complexos em elevada dimensão.

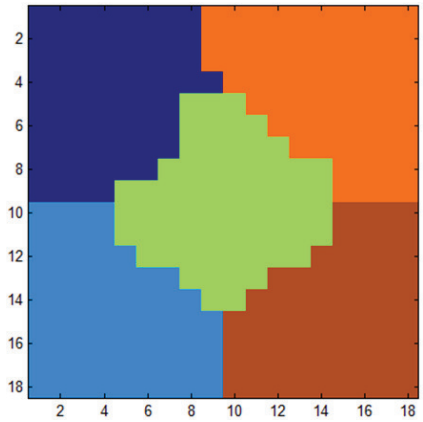

Figura 8: regiões conectadas de neurônios com uso da GC-*matrix*.

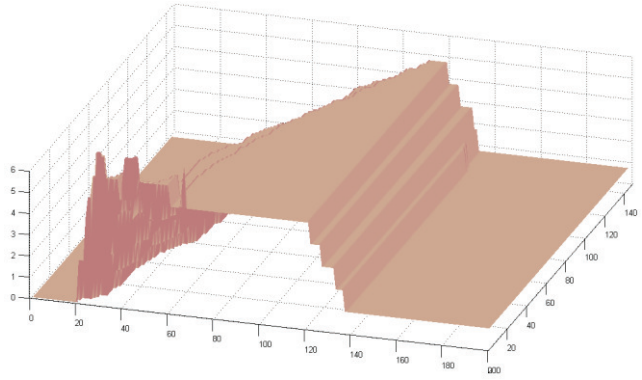

Figura 9: Número de agrupamentos, $N_{rc}^{\ k}$ (eixo z) versus k, limiar de distância (em x), e t, parâmetro de área do filtro morfológico (no eixo y), para o SOM com tamanho 18x18, vizinhança final de treinamento 1, com uso da *U-matrix*.

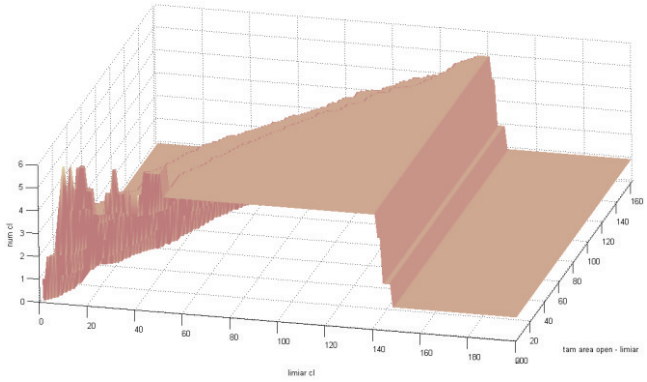

Figura 10: Número de agrupamentos, $N_{rc}^{\ k}$ (eixo z) versus k, limiar de distância (em x), e t, parâmetro de área do filtro morfológico (no eixo y), para o SOM com tamanho 18x18, vizinhança final de treinamento 1, com uso da *GC-matrix*.

5.2 O conjunto de dados *chainlink*

Um exemplo não trivial (não linearmente separável) para comparações de métodos de agrupamentos em dados multidimensionais foi proposto por Ultsch (1995), que é o *chainlink*, consistindo de 1000 pontos no espaço $\Re^3$ de forma

similar a dois anéis tridimensionais entrelaçados. Um dos anéis se estende na direção *x-y* enquanto o outro se estende na direção de *x-z*. Os dois anéis podem ser pensados como elementos de uma corrente, cada um consistindo de 500 objetos de dados.

Este problema ilustra a capacidades do SOM em descobrir a estrutura dos dados mesmo para conjuntos de dados com forma complexas e não-esféricas, e não separáveis linearmente. Alguns destes dados foram apresentados em Costa e Netto (1999b).

Foi utilizado um mapa com tamanho 15 × 15. A inicialização de pesos foi linear e o SOM foi treinado com o algoritmo de atualização em lote (*batch*). A função de vizinhança usada foi Gaussiana e o raio inicial foi 12, caindo para 1 de forma linear com o tempo. O número de épocas foi fixado em 500. A Figura 11 ilustra a configuração dos neurônios no espaço 3-D após o final do treinamento. A relação de vizinhança é expressa por linhas que conectam os neurônios.

A Figura 12 ilustra, em pseudo cor, a *U-matrix* (na esquerda) e *GC-matrix* (direita), obtidas para o SOM com tamanho 15x15, a partir do conjunto de dados *Chainlink*. A mesma informação é apresentada, em forma de superfície, nas Figuras 13a 13b. De forma similar ao comentário da Seção 5.1, nota-se que a *GC-matrix* apresenta maior suavidade, facilitando não apenas a visualização dos agrupamentos mas também a segmentação. A Figura 14 apresenta linhas da *watershed* sobrepostas na *U-matrix.* Em ambos os casos houve a descoberta do número correto de agrupamentos (2) e a recuperação da estrutura dos dados, obtendo acurácia de 100%. A Figura 15 mostra a segmentação da *U-matrix* onde os dois agrupamentos são mostrados, separados pelas linhas de *watershed* (em preto). Métodos como os apresentados em Vesanto e Alhoniemi (2000) e Brugger et al. (2008) não foram capazes de detectar o número correto de classes para o *chainlink*.

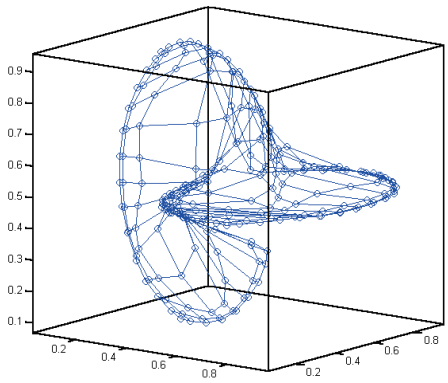

Figura 11 - Grid do SOM 15×15 após 500 iterações usando o algoritmo de atualização em lote

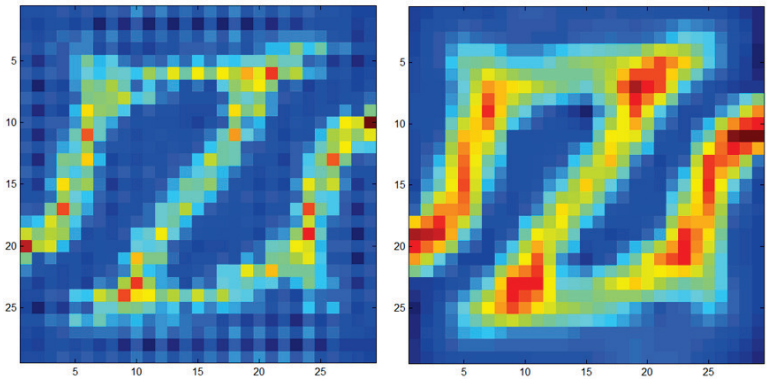

Figura 12: (a) esquerda, *U-matrix*; e (b) direita, *GC-matrix*, obtidas para o SOM com tamanho 15x15, a partir do conjunto de dados Chainlink.

© Sociedade Brasileira de Redes Neurais (SBRN)

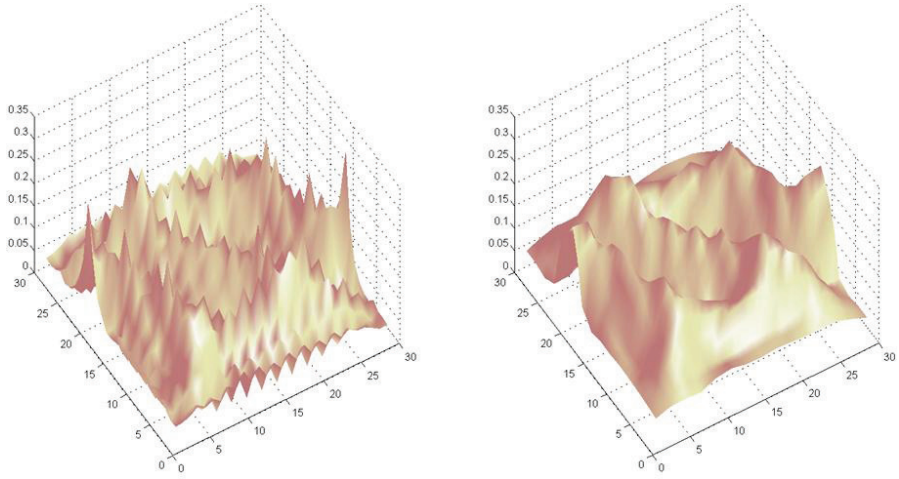

Figura 13: (a) *U-matrix (*esquerda) e (b) *GC-matrix* (direita), em forma de superfície, obtidas para o SOM com tamanho 15x15, a partir do conjunto de dados Chainlink.

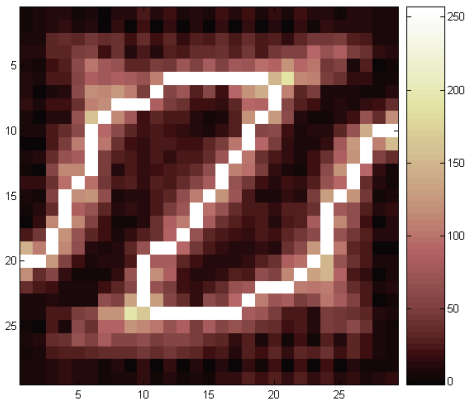

Figura 14: linhas da *watershed* sobrepostas na *U-matrix*

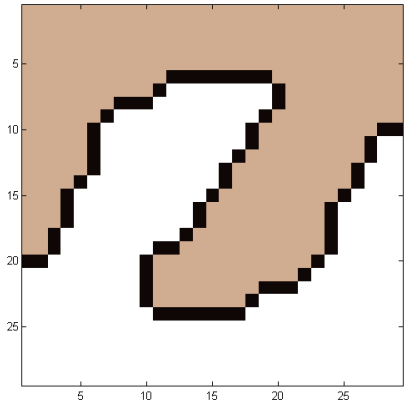

Figura 15: Partição da *U-matrix* (já rotulada) onde os dois agrupamentos são mostrados separados pelas linhas de watershed (em preto).

6. Conclusões

Dado o volume e complexidade das bases de dados nas mais diversas aplicações atuais, dispor de ferramentas que possam processar de forma automática grandes massas de dados não rotuladas é de grande importância.

Visualização e agrupamentos estão entre principais tarefas de mineração de dados. Mapas tipo SOM têm sido bastante utilizados em várias aplicações, porém, etapas posteriores ao treinamento devem ser adicionadas para extrair o conhecimento obtido a partir da configuração de pesos dos neurônios.

Neste aspecto, o uso do SOM gerando imagens a partir de dados multivariados e seu processamento, obtendo por segmentação regiões de neurônios associados aos *clusters* dos dados em elevada dimensão, habilita o melhor entendimento da estrutura de dados, permitindo diversas outras ações de apoio a decisão.

Em visão computacional, diversos algoritmos são utilizados para processar, nos mais variados níveis, e tentar entender, de forma automática ou semi-automática, aspectos de uma dada imagem (ou conjunto de imagens). A contribuição deste artigo está na proposta de uma nova forma de visualização da rede neural SOM, a *GC-matrix*, e sua comparação com a forma tradicional, a *U-matrix*.

Resultados são apresentados tanto no aspecto de visualização quanto em agrupamentos de dados, para mapas de tamanhos diferentes e vizinhanças finais diferentes. Bases de dados derivadas de misturas de Gaussianas foram testadas, assim como a base de dados *Chainlink*, que apresenta desafios na área de agrupamentos, devido ao caráter não linearmente separável e formato complexo.

Foi mostrado que a *GC-matrix* apresentou uma maior suavidade, e isso implicou em uma maior estabilidade para escolha de parâmetros importantes, como o limiar de área para suavização ou filtragem morfológica, etapa anterior a escolha de marcadores para a *watershed*, determinante do número de *clusters*. Resultados com bases de dados reais, como Iris e Wine (Asuncion e Newman, 2007) estão de acordo com os encontrados para a Seção 5.

Trabalhos futuros podem explorar o uso da *GC-matrix* com índices de validação, como o CDbw (Halkidi e Vazirgiannis, 2008) e comparações com diferentes imagens (superfícies), como a descrita em (Costa et al., 2011).

7. Agradecimento

Agradecemos o apoio financeiro do CNPq e aos comentários do Prof. Hujun Yin (University of Manchester) e dos revisores.

8. Referências

Asuncion, A. e Newman, D.J. (2007). *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science. URL [http://www.ics.uci.edu/~mlearn/MLRepository.html]

Brugger, D., Bogdan, M. e Rosenstiel, W. (2008). Automatic Cluster Detection in Kohonen's SOM. *IEEE Transactions on Neural Networks*, Vol. 19, No. 3, pp. 442-459.

Costa, J.A.F. e Netto, M.L.A. (1999). "Estimating the Number of Clusters in Multivariate Data by Self-Organizing Maps". *Intl. Journal of Neural Systems*, vol. 9, pp. 195-202.

Costa, J.A.F. e Netto, M.L.A. (2001). "Clustering of complex shaped data sets via Kohonen maps and mathematical morphology". In: *Proceedings of the SPIE, Data Mining and Knowledge Discovery.* B. Dasarathy (Ed.), Vol. 4384, pp. 16-27.

Costa, J.A.F. e Netto, M. L. A. (1999). Cluster Analysis Using Self-Organizing Maps and Image Processing Techniques. In: *Proc. of the 1999 IEEE International Conference on Systems, Man, and Cybernetics(SMC'99),* vol. 5, pp. 367-372, Tokyo, Japan.

Costa, J.A.F. e Netto, M.L.A. (2007). Segmentação de Mapas Auto-Organizáveis com Espaço de Saída 3-D. *Controle & Automação* - Ed. Especial Automação Inteligente. Vol.18 no.2, 2007, pp. 150-162.

Costa, J.A.F., Gonçalves, M. L. e Netto, M.L.A (2011). Visualização e análise de agrupamentos usando redes auto-organizáveis, segmentação de imagens e índices de validação. Learning and NonLinear Models. Submetido.

Dougherty, E. R. e Lotufo, R. A. (2003). *Hands-on Morphological Image Processing*. SPIE Publications.

Halkidi, M. e Vazirgiannis, M. (2008). A Density-based Cluster Validity Approach using Multi-representatives, *Pattern Recognition Letters*, Vol. 29, pp. 773-786.

Hamad, D., Firmin, C. e Postaire, J. (1996). Unsupervised pattern classification by neural networks. *Mathematics and Computers in Simulation*, v. 41, pp. 109-116.

Kaski, S., Nikkilä, J. e Kohonen, T., (2000). Methods for exploratory cluster analysis. In: *Proceedings of SSGRR 2000, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, L'Aquila, July 31--August 6. Scuola Superiore G. Reiss Romoli, 2000.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd Ed., Springer Verlag, Berlin.

SOM Toolbox 2.0, URL: http://www.cis.hut.fi/projects/somtoolbox/.

Ultsch, A. (1993). "Self-Organizing Neural Networks for Visualization and Classification". In: O. Opitz et al. (Eds). *Information and Classification*, pp.301-306. Springer: Berlin..

Ultsch, A. (1995). Self-Organizing Neural Networks perform different from statistical k-means clustering. *Gesellschaft für Klassifikation*, Basel.

Vesanto, J. e Alhoniemi, E. (2000). Clustering of the Self-Organizing Map, *IEEE Trans. on Neural Networks*, 11, (3), pp. 586-602.

Vesanto, J. (2000). *Using SOM in Data Mining.* Licentiate's Thesis, Department of Computer Science and Engineering, Helsinki University of Technology, Espoo, Finland.

Witten, I. H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed., Morgan Kaufmann.

Xu, R. e Wunsch, D. (2009). *Clustering*, IEEE Press.