Segmentação Semi-Automática de Dados Geo-espaciais Multivariados com Mapas Auto-Organizáveis

Marcos A.S. da Silva[1,2] , Antônio M.V. Monteiro[1] , José S. de Medeiros[1]
[1]*Instituto Nacional de Pesquisas Espaciais*
Divisão de Processamento de Imagens
São José dos Campos, SP
{miguel,simeao} @dpi.inpe.br
[2]*Laboratório de Geotecnologias Aplicadas*
Embrapa Tabuleiros Costeiros, Aracaju, SE
aurelio@cpatc.embrapa.br

Resumo

Este trabalho avaliou o uso do algoritmo de segmentação automática do Mapa Auto-Organizável (SOM), Costa-Neto, em conjunto com os índices de validação de partição de dados, Davies-Bouldin e CDbw na descoberta de agrupamentos em dados geo-espaciais multivariados.

Abstract

This paper evaluated the use of the automatic segmentation algorithm of the Self-organizing Map (SOM), Costa-Neto, together with clustering validity indexes, Davies-Bouldin and CDBw, in the clustering task in multivariate geospatial data.

1. Introdução e Motivação

Existem vários mecanismos de análise exploratória de dados através dos mapas de Kohonen [6]. Porém, a tarefa de descoberta de agrupamentos tem sido feita visualmente, através da projeção do mapa por meio da U-matriz [7] e dos planos de componentes [6]. Todavia, existem casos onde a complexidade da U-matriz gerada inviabiliza ou dificulta a descoberta de agrupamentos pela verificação visual. Para estes casos seria bastante útil a existência de técnicas de detecção automática de agrupamentos baseados nos vetores de código gerados pelo SOM.

Este trabalho avaliou o método de segmentação automática do SOM, proposto por [2]. Este método foi aplicado em conjunto com os índices de validação de partição de dados, índice Davies-Bouldin [3] e o CDbw [5].

A seção 2 faz uma breve introdução sobre os Mapas Auto-Organizáveis. Na seção 3 o algoritmo Costa-Neto é detalhado e uma rápida comparação com outros métodos de segmentação automática do SOM é realizada. A seção 4 descreve os algoritmos de cálculo dos índices Davies-Bouldin e CDbw. As seções 5, 6, 7 e 8 tratam da metodologia, estudo de caso, resultados e conclusões, respectivamente.

2. Mapa Auto-Organizável

O Mapa Auto-Organizável de Kohonen é uma rede neural de aprendizagem competitiva organizada em duas camadas [6]. A primeira camada representa o vetor dos dados de entrada, x_k, a segunda corresponde a uma grade de neurônios, geralmente bidimensional, totalmente conectada aos componentes do vetor de entrada. Cada neurônio possui um vetor de código associado, w_j.

O processo de aprendizagem consiste de três fases. Na primeira fase, competitiva, cada padrão de entrada é apresentado a todos os neurônios para que aquele mais próximo do padrão apresentado seja o vencedor. Na segunda fase, cooperativa, é definida a vizinhança relativa ao neurônio vencedor. Na terceira fase, adaptativa, os vetores de código do neurônio vencedor e dos seus vizinhos serão alterados segundo algum critério de atualização.

Após o processo de aprendizagem os vetores de código do SOM corresponderam a uma aproximação não-linear dos padrões de entrada. O SOM também preserva a formação topológica dos padrões, ou seja, padrões próximos no conjunto amostral estarão relacionados a neurônios próximos na grade neural.

O SOM pode variar quanto a algoritmos de aprendizagem, estrutura topológica da grade, função de vizinhança, parametrização inicial etc.

3. Segmentação Automática do SOM

Para a tarefa de segmentação da grade de neurônios do SOM após a fase de treinamento analisou-se os métodos Costa[1], Vesanto [8] e Costa & Netto [2]. Em função de ser um método conceitualmente simples, baseado nas informações contidas, unicamente, nos neurônios e na rede após o treinamento, aplicável a Mapas com diferentes topologias e totalmente automático optou-se pelo algoritmo Costa-Neto [2] para segmentar o SOM treinado.

3.1 Algoritmo de Segmentação Costa-Neto

Considerando a camada de saída da rede como uma estrutura de grafo não orientado e conectado segundo a estrutura de vizinhança do SOM. Costa & Netto [2] propõe a segmentação do mapa baseado no particionamento deste grafo. O algoritmo Costa-Neto visa eliminar conexões inconsistentes entre os neurônios para encontrar a partição ideal da grade.

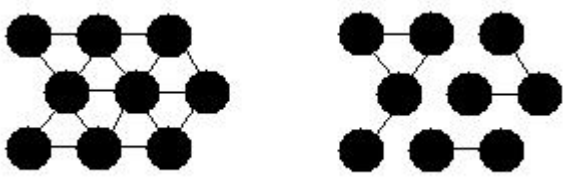

Figura 1. **O algoritmo Costa-Neto elimina conexões inconsistentes entre os neurônios para encontrar a partição ideal da grade.**

O algoritmo proposto baseia-se em informações geométricas de distância entre os neurônios, erro de quantização e atividade do neurônio. A estratégia é considerar que todos os neurônios fazem parte de um grafo não orientado parcialmente conectado e, a partir de regras heurísticas, eliminar conexões inconsistentes entre neurônios vizinhos, Figura 1.

O Algoritmo:

1.1. Obter as distâncias entre os pesos dos neurônios adjacentes *i* e *j*, $d(w_i,w_j)$; e a atividade de cada neurônio *i*, *H(i)*.
1.2. Para cada par de neurônios adjacentes, *i* e *j*, a aresta será considerada inconsistente caso:
  1.2.1. Se a distância entre os pesos excede em 2 a distância média dos outros neurônios adjacentes a i ou a j;
  1.2.2. Se os dois neurônios adjacentes *i* e *j* possuem atividade *(H)* abaixo de 50% do mínimo permitido (H_{min}), ou um dos neurônios for inativo (*H(i) = 0*);
  1.2.3. Se a distância entre os centróides dos conjuntos de dados associados aos neurônios *i* e *j* exceder em 2 vezes a distância entre os pesos $d(w_i, w_j)$.
1.3. Remoção dos ramos (arestas) inconsistentes. Para cada aresta *(i, j)* considerada inconsistente resultará em uma conexão nula no endereço *(i, j)* da matriz de adjacência *A*. Ramos consistentes recebem entrada 1 no endereço *(i, j)* de *A*.
1.4. Atribuir um código distinto para cada conjunto de neurônios conectados.
1.5. Remover grupos conectados pequenos (com menos de 3 neurônios).

O que acontece com a aplicação do algoritmo é uma poda dos neurônios conectados adjacentes. Ou seja, ao final têm-se vários grupos de neurônios conectados representando um agrupamento específico.

O algoritmo proposto por [2] é independente da U-matriz e da dimensionalidade da grade do Mapa, o que o torna mais genérico que a proposta de segmentação baseada na U-matriz [1].

O algoritmo faz uso de alguns limiares empíricos definidos por meio de experimentações, e consegue particionar os dados usando somente as informações inerentes ao Mapa treinado, como a distância entre os neurônios, erro de quantização e nível de atividade.

O algoritmo de detecção automática de agrupamentos baseado na partição do SOM [2] separa os padrões mas não garante que todos os vetores de entrada serão rotulados. Por exemplo, dados atípicos podem não ser rotulados devido a restrição 1.5 do algoritmo seção 3.1.

Este problema pode ser solucionado usando-se o critério do vizinho mais próximo para rotulação de todos os neurônios especializados do Mapa. Este procedimento evitará que os cálculos dos índices de validação sejam comprometidos.

4. Validação da Partição dos Dados

Para validar os agrupamentos gerados pelo algoritmo de segmentação baseado no particionamento de grafos foram usados dois índices, já aplicados aos Mapas Auto-Organizáveis. O índice Davies-Bouldin [3], foi usado em [8] para auxiliar o processo de definição do número de agrupamentos corretos. O índice CDbw [5] foi usado em [9] numa aplicação semelhante a anterior.

4.1 Índice Davies-Bouldin

O índice Davies-Bouldin [3] é uma medida que indica a similaridade entre agrupamentos. Esta medida pode ser

usada para avaliação da partição dos dados e, conseqüentemente, para comparação relativa entre diferentes divisões do conjunto de dados. O índice Davies-Bouldin é independente do número de agrupamentos e do método de partição dos dados, o que o torna indicado para avaliação de algoritmos de partição de dados.

O índice Davies-Bouldin é dado por

$$\frac{1}{c}\sum_{k=1}^{c}\max_{c\neq k}\left\{\frac{S_c(Q_k)+S_c(Q_l)}{d_{ce}(Q_k,Q_l)}\right\}$$

onde c é o número de agrupamentos, $S_c(Q_k)$ representa a distância intra-agrupamento (dispersão interna do agrupamento Q_k), baseado na distância para o centróide, $d_{ce}(Q_k,Q_l)$ representa a distância entre os agrupamentos Q_k e Q_l, também baseado na distância entre os centróides. $S_c(Q_k)$ é dado por

$$S_c(Q_k)=\left\{\frac{1}{N_K}\sum_{j=1}^{N_k}\left|x_j-c_k\right|^q\right\}^{1/q}$$

onde $x_j \in Q_k$, N_k é o número de amostras no agrupamento Q_k e $c_k=1/N_k \Sigma_{xi \in Qk} x_i$. d_{ce} é dado por

$$d_{ce}(Q_k,Q_l)=\left\{\sum_{k=1}^{d}\left|c_{ki}-c_{kj}\right|^p\right\}^{1/p}$$

onde d corresponde a dimensão do vetor x_k.

No índice Davies-Bouldin uma boa partição dos dados é indicada para valores baixos. Os valores p e q devem ser escolhidos convenientemente de acordo com o problema.

Ultsch [7] usou o índice Davies-Bouldin com p=2 e q=2 para avaliação da partição dos dados feita através do Mapa Auto-Organizável. Neste caso o SOM foi usado como um redutor do volume de dados a ser particionado. Após esta redução dois métodos de partição tradicionais, o k-médias e o método hierárquico aglomerativo, foram aplicados, separadamente, para encontrar os agrupamentos. O índice Davies-Bouldin foi usado para atuar como critério de junção ou separação de agrupamentos nos algoritmos de partição usados.

Uma das principais características deste índice é sua adequabilidade para estruturas hiperesféricas, já que o mesmo usa o centróide como ponto de referência.

4.2 Índice CDbw

O índice CDbw (*Compose Density between and within clusters*) [5] também basea-se na medição das distâncias intra e inter agrupamentos. Porém, enfatiza as características geométricas de cada agrupamento, tratando eficientemente agrupamentos com formatos arbitrários.

A característica geométrica do agrupamento é representada através do uso de vetores representativos de cada agrupamento. Ao invés de usar o centróide como referência, usa-se um conjunto de vetores. Isto permite que o índice avalie corretamente estruturas não hiperesféricas, o que não ocorre com o índice Davies-Bouldin.

Para um conjunto de dados particionados em c agrupamentos, define-se um conjunto de pontos representativos $V_i=\{v_{i1},v_{i2},...,v_{iri}\}$ para o agrupamento i, onde r_i representa o número de pontos de representação para o agrupamento i.

Para cada componente ρ do agrupamento i tem-se que o desvio padrão $stdev(i)$ é dado por

$$stdev^{\rho}(i)=\sqrt{\sum_{k=1}^{n_i}(x_k^{\rho}-m_i^{\rho})^2/(n_i-1)} \quad (1)$$

onde n_i representa o número de amostras no agrupamento i, $x_k \in Q_i$, e m_i a média da amostra do i-ésimo agrupamento. A média do desvio padrão é dada por

$$stdev=\frac{1}{c}\sqrt{\sum_{i=1}^{c}\left\|stdev(i)\right\|^2} \quad (2)$$

A densidade intra-agrupamento é definida como

$$Intra\_dens(c)=\frac{1}{c}\sum_{i=1}^{c}\frac{1}{r_i}\sum_{j=1,j\neq i}^{r_i}density(v_{ij}),c>1 \quad (3)$$

O termo $density(v_{ij})$ é definido como $\text{density}(v_{ij})=\sum_{l=1}^{n_i} f(x_l,v_{ij})$, onde $x_l \in Q_i$, v_{ij} é a j-ésima representação do i-ésimo agrupamento, e $f(x_l,v_{ij})$ é dado por

$$f(x_l,v_{ij})=\begin{cases}1, & \left\|x_l-v_{ij}\right\| \leq stdev \\ 0, & \text{caso contrário}\end{cases} \quad (4)$$

A densidade inter-agrupamento é dada por

$$Inter\_dens(c) = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} \frac{\left\| close\_rep(i) - close\_rep(j) \right\|}{\left\| stdev(i) \right\| + \left\| stdev(j) \right\|} density(v_{ij}), c > 1, c \neq n \quad (5)$$

onde *close_rep(i)* e *close_rep(j)* representam o par de pontos de representação mais próximos entre o agrupamento *i* e o *j*, v_{ij} é o ponto médio entre este par de pontos. *density*(v_{ij}) é dado por density(v_{ij})= $\sum_{k=1}^{n_i+n_j} f(x_k, v_{ij})$, onde $x_k \in Q_i$ ou $x_k \in Q_j$, e $f(x_k, v_{ij})$ é dado por

$$f(x_k, v_{ij}) = \begin{cases} 1, \left\| x_k - v_{ij} \right\| \leq (\left\| stdev(i) \right\| + \left\| stdev(j) \right\|) \\ 0, \quad \text{caso contrário} \end{cases} \quad (6)$$

A separação entre os agrupamentos é dada por

$$Sep(c) = \sum_{i=1}^{c} \sum_{j=1, j\neq i}^{c} \frac{\left\| close\_rep(i) - close\_rep(j) \right\|}{1 + Inter\_dens(c)}, c > 1 \quad (7)$$

O índice CDbw é definido por

$$CDbw(c) = Intra\_dens(c) * Sep(c) \quad (8)$$

Uma boa partição dos dados é indicada para valores altos do índice. A complexidade *O(n)* do algoritmo [5] é favorável para dados geoespaciais.

Uma questão importante a ser considerada neste algoritmo é a definição dos vetores de referência para cada agrupamento. Segundo Halkidi & Vazirgiannis [5] este processo é iterativo, primeiro escolhe-se o ponto mais distante da média do agrupamento, posteriormente o ponto mais distante do ponto anterior é escolhido e assim sucessivamente.

4.3 Usando os vetores de código como vetores de referência no CDbw

Como colocado por [5] os vetores de referência, para o cálculo do CDbw, pode ser encontrado de forma iterativa a partir do conjunto de dados particionado. Todavia, em [5] não há uma definição do critério de parada para o algoritmo de criação dos vetores de referência. Ou seja, o número de vetores de referência, para cada agrupamento, tem de ser definido empiricamente para servir como critério de parada, caso contrário todos os vetores poderiam ser escolhidos como vetores de referência.

Para o caso de partição dos dados através do SOM os vetores de códigos funcionam como uma aproximação não-linear da distribuição dos dados de entrada sendo, portanto, vetores representativos dos dados amostrais. Logo, pode-se usar os vetores de código do SOM particionado como vetores de referência dos seus respectivos agrupamentos. Isto simplifica o processo de cálculo do CDbw para o caso de partição dos dados com SOM.

A adequação desta abordagem dependerá da relação entre o número *n* de padrões e o número *m* de neurônios. Para *m/n* muito pequeno pode-se ter uma deficiência em número de neurônios para representação de cada agrupamento. Para *m/n* muito grande tem-se o inverso.

5. Metodologia

O particionamento do conjunto de dados num número c de agrupamentos foi realizado através do algoritmo Costa-Neto em duas fases, Figura 2. Primeiramente os dados foram apresentados ao SOM, este é treinado e então seus vetores de código particionados. Como cada padrão está associado a um vetor de código pode-se particionar os dados a partir dos vetores de código particionados.

Para validação dos agrupamentos gerados usou-se o índice Davies-Bouldin, (p=2,q=1);(p=2,q=2), e o Cdbw.

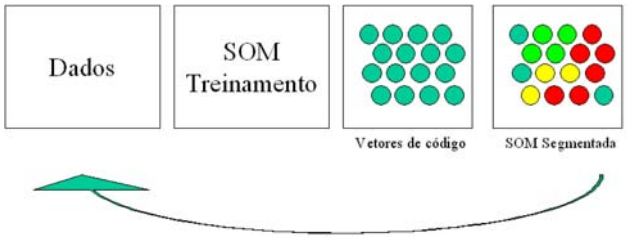

Figura 2. **Fases do processo de particionamento dos dados em *c* agrupamentos**

O processo de avaliação dos agrupamentos usado neste trabalho seguiu o seguinte roteiro.

1. Definiu-se um conjunto de redes que foram testadas e para cada rede:
   a. Efetuou-se o treinamento da rede;
   b. Aplicou-se o algoritmo de detecção de agrupamentos Costa-Neto;
   c. Rotulou-se todos os neurônios através do método do vizinho mais próximo;
   d. Calculou-se os índices Davies-Bouldin e CDbw;
2. Escolheu-se a rede com os melhores valores dos índices.

6. Estudo de Caso

A análise de exclusão/inclusão social urbana em São José dos Campos-SP foi baseada nos estudos conduzidos por Genovez em sua dissertação de mestrado [4].

A metodologia consiste na análise de atributos associados aos setores censitários da área urbana de São José dos Campos. Cada setor censitário possui um conjunto de atributos relativos aos dados do IBGE que correspondem a questões relacionadas com o nível de qualidade de vida daquela população.

Para este trabalho foram usados 8 índices de exclusão/inclusão social urbana. Estes índices são composições de dados brutos relativos aos setores censitários. Cada índice varia entre –1, maior nível de exclusão social, e +1, maior nível de inclusão social.

Os 8 índices correspondem ao nível de renda familiar, desenvolvimento educacional, estímulo educacional, longevidade, qualidade ambiental, concentração de mulheres chefes de família e mulheres não alfabetizadas.

7. Resultados

Do gráfico correspondente ao índice Davies-Bouldin, Figura 3, tem-se que a melhor partição é a da rede 14x10, com índices Davies-Bouldin 3.0 e 1.5 e c=3. Porém, ao analisar o SOM rotulado, Figura 4, percebe-se que o particionamento não corresponde a realidade, uma vez que coloca no mesmo grupo neurônios especializados em setores de exclusão e inclusão.

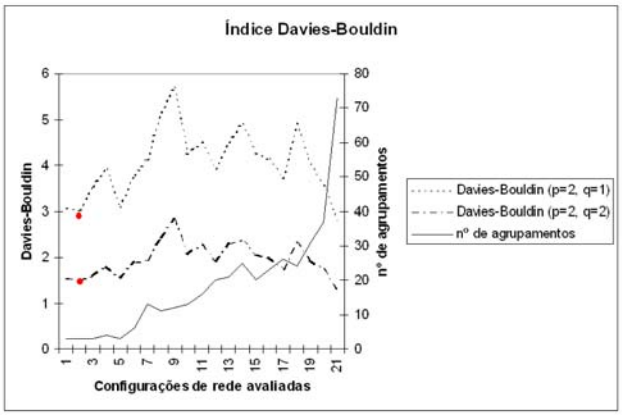

Figura 3. **Gráfico para o índice Davies-Bouldin**

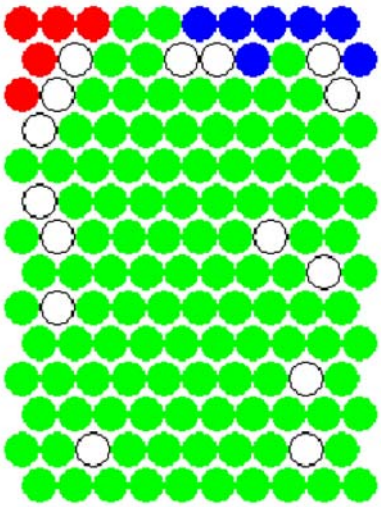

Figura 4. **Mapa neural particionado segundo o índice Davies-Bouldin**

Do gráfico correspondente ao índice CDbw, Figura 5, tem-se que a melhor partição é a da rede 18x16, com índice CDbw igual a 110,14 e c=20. Da análise do Mapa particionado, Figura 6, conclui-se que a partição obedece ao sentido da distribuição vertical do Mapa e que identifica claramente as zonas de dados atípicos. O mapa dos setores censitários da cidade de São José dos Campos foi colorido segundo esta partição do SOM, Figura 7, e demonstra coerência com os resultados obtidos por [4], no sentido de identificação de áreas de inclusão e exclusão social urbana.

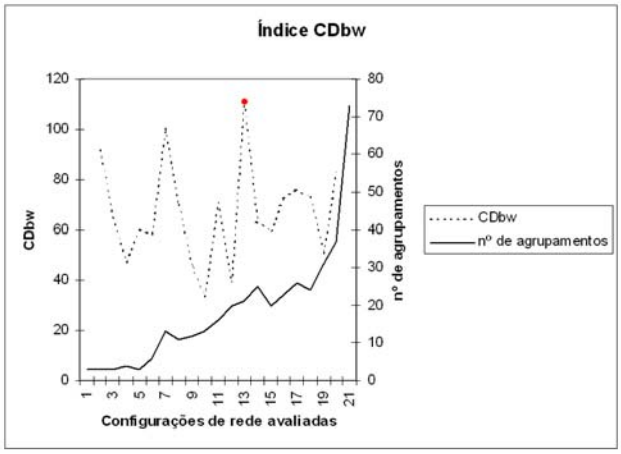

Figura 5. **Gráfico para o índice CDbw**

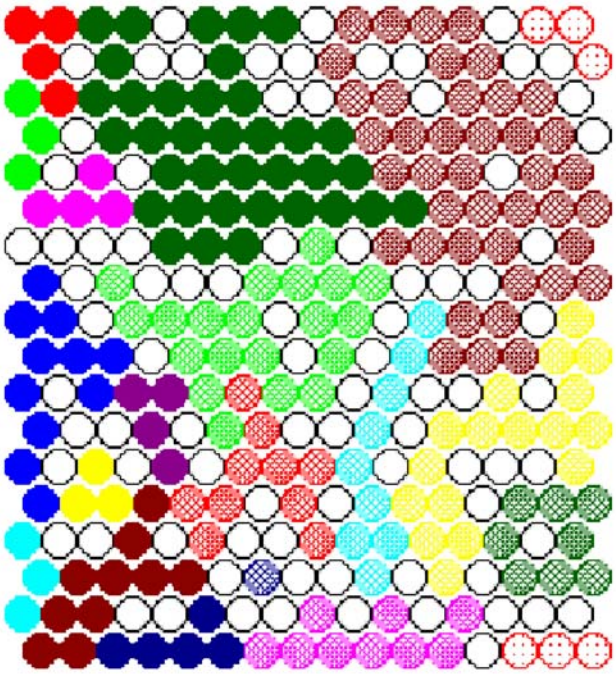

Figura 6. **Mapa particionado segundo o índice Cdbw**

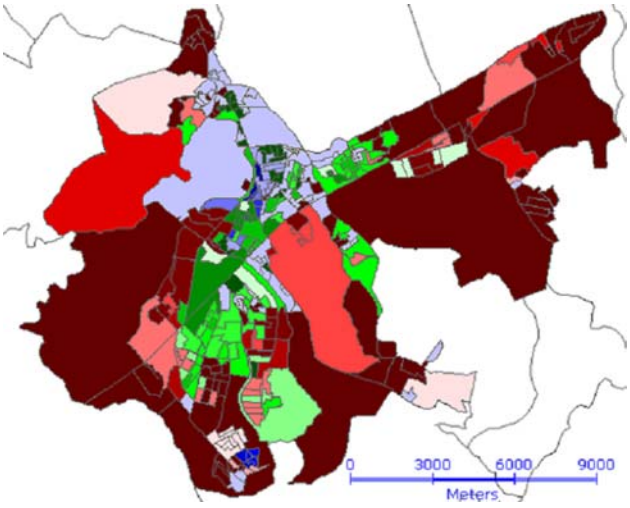

Figura 7. **Mapa coroplético dos setores censitários gerados a partir do SOM particionado segundo o algoritmo Costa-Neto e validação do índice CDbw**

8. Conclusões

Para o estudo de caso em questão a combinação do algoritmo Costa-Neto com o índice de validação CDBw mostrou-se coerente com resultados anteriores para o mesmo conjunto de dados. Ou seja, a partição dos dados segundo esses algoritmos identificou e agrupou em zonas de exclusão ou inclusão social urbana.

Embora o índice CDbw tenha se saído melhor neste caso mais estudos devem ser conduzidos a fim de se avaliar os índices para outros dados com estruturas topológicas distintas.

O processo não chega a ser automático em função da necessidade do usuário definir as dimensões da rede e o raio inicial para cada configuração de rede a ser avaliada. Todavia, pode-se criar um algoritmo simples para automatizar esta fase e tornar o processo totalmente automático.

9. Referências

[1] Costa, J. A. F. **Classificação automática e análise de dados por redes neurais auto-organizáveis**. São Paulo. Tese – Faculdade de Engenharia Elétrica e de Computa¸c˜ao - UNICAMP, Dezembro 1999.

[2] Costa, J. A. F.; Netto, M. L. A. Segmentação do SOM Baseada em Particionamento de Grafos. Congresso Brasileiro de Redes Neurais. **Anais do VI Congresso Brasileiro de Redes Neurais**. 2003.

[3] Davies, D. L.; Bouldin, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, p. 224–227, 1979.

[4] Genovez, P. C. **Território e Desigualdades: Análise Espacial Intra-urbana no estudo da dinâmica de exclusão/inclusão social no espaço urbano em São José dos Campos-SP**. Dissertação – INPE, Dezembro 2002.

[5] Halkidi, M.; Vazirgiannis, M. Clustering validity assessment using multi representatives. **Proceeding of SETN Conference**, Thessaloniki, Greece. 2002.

[6] Kohonen, T. **Self-organizing maps**. Springer, 1995. Third Edition 2001.

[7] Ultsch, A. **Information and Classification**. Springer, 1993. Cap. Knowledge extraction from self-organizing neural networks.

[8] Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks, v. 11, n. 3, p. 586–600, May 2000.

[9] Wu, S.; Chow, T. W. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. **Pattern Recognition**, 2003. In Press.