Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, Vol. 9, Iss. 2, pp. 91-103, 2011

VISUALIZAÇÃO E ANÁLISE DE AGRUPAMENTOS USANDO REDES AUTO-ORGANIZÁVEIS, SEGMENTAÇÃO DE IMAGENS E ÍNDICES DE VALIDAÇÃO

José Alfredo F. Costa[1], Márcio L. Gonçalves[2] e Márcio L. de Andrade Netto[3]

[1]Departamento de Engenharia Elétrica, Universidade Federal do Rio Grande do Norte (UFRN) – Natal, RN
[2]Curso de Ciência da Computação - PUC Minas, Poços de Caldas, MG
[3]Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas – Campinas, SP
E-mails: alfredo@ufrnet.br, marcio@pucpcaldas.br, marcio@dca.fee.unicamp.br

Resumo- Mapas auto-organizáveis (SOM) têm sido amplamente utilizados em agrupamento (quantização) e visualização de dados multivariados, uma importante propriedade que não existe na maioria dos algoritmos de agrupamentos. Porém, análise de aglomerados efetiva utilizando o SOM envolve duas ou três etapas do processamento. Após o treinamento, neurônios são agrupados gerando regiões no mapa que estão relacionados aos aglomerados do conjunto de dados. O pressuposto básico depende da aproximação da densidade de dados, obtida através de aprendizagem não supervisionada. A transformação de dados em imagens, sua visualização e análise automática de agrupamentos são abordadas neste artigo. O método proposto segmenta o mapa através da transformada watershed, após modificação da homotopia da imagem, e utilizando índices de validação de agrupamentos. Os resultados são mostrados para conjuntos de dados para mapas com diferentes tamanhos.

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

Abstract- Self-organizing maps (SOM) had been widely used for input data quantization and visual display of data, an important property that does not exist in most of clustering algorithms. Effective data clustering using SOM involves two or three steps procedure. After proper network training, units can be clustered generating regions of neurons which are related to data clusters. The basic assumption relies on the data density approximation by the neurons through unsupervised learning. The transformation of high dimensional data into images and its visualization and clustering is addressed in this paper. The proposed method segments SOM networks via watershed algorithms and modified cluster validation indexes. Results are shown for benchmark datasets for different map sizes.

Keywords- Self-organizing maps, visualization, data clustering, validation indexes, 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.

Mineração de dados, parte do processo de descoberta de conhecimento em bases de dados, tem obtido bastante atenção dada o aumento e complexidade das bases de dados nas mais diversas aplicações (Witten & Frank, 2005). Uma série de processos foi desenvolvida para lidar com dados multivariados. Métodos supervisionados são usados quando há disponibilidade das saídas desejadas para um conjunto de entradas disponíveis. Neste caso, um sinal de erro ou discrepância entre a saída desejada e a obtida pelo sistema são utilizadas no algoritmo de aprendizado para otimização dos parâmetros internos e minimizar futuros erros. Exemplos de tais métodos incluem regras de classificação e predição.

No entanto, em muitos casos, o custo de rotulagem dos dados pode ser elevado ou mesmo impossível. Neste caso, os métodos não supervisionados, que trabalham diretamente com os dados de entrada, são aplicados. Exemplos de técnicas neste paradigma incluem métodos de redução de dimensionalidade e de agrupamento de dados (*clustering*). As técnicas de agrupamento têm obtido bastante destaque na área de mineração de dados, tanto em pesquisas quanto aplicações (Xu & Wunsch, 2009). O objetivo é encontrar uma organização conveniente e válida de dados multivariados, com base nas semelhanças entre os padrões. As técnicas de agrupamento podem ser classificados, de forma geral, em cinco grandes classes

(Jain et al., 1999): métodos de particionamento; métodos hierárquicos; métodos baseados em estimação de densidades; algoritmos baseados em grafos; e métodos baseados em misturas de modelos de densidades. Em muitos casos, é possível derivar regras e perfil geral dos grupos, que podem auxiliar na classificação de novos dados não rotulados.

A segmentação de imagens, um importante problema em visão computacional, pode ser formulada como um problema de agrupamento. Documentos também podem ser agrupados para gerar tópicos hierárquicos para o acesso ou recuperação da informação. Exemplos incluem ainda segmentação de mercado (por exemplo, em marketing) e no monitoramento de processos, no qual um determinado estado da planta corresponde a um cluster. Aplicações de análise de agrupamentos aparecem nas mais diversas áreas, como análise de preferência de alimentos (Sahmer et al., 2006), redes de sensores (Younis e Fahmy, 2004), compressão de sinais e imagens (Xu & Wunsch, 2005), na biologia e medicina (Everitt et al., 2001; Gan et al. 2008; Xu and Wunsch, 2009), entre várias outras.

Boas referências introdutórias incluem Gan et al (2007), Xu e Wunsch (2005, 2009), Jain et al. (1999) e Kogan (2006), além das referências históricas importantes, como Anderberg (1973), Hartigan (1975), Jain e Dubes (1988), Kaufman e Rousseeuw (1990), Arabie et al. (1999) e Everitt et al. (2001).

Entretanto, a maioria dos métodos de agrupamento de dados possuem geometrias pré-estabelecidas (inerentes, dada a escolha do critério de similaridade e fusão de dados e classes) que impõem uma estrutura aos dados, ao contrário de tentar descobrir. Exemplos incluem o k-means (hiper-esferas, no caso do uso de distância Euclidiana) e outros mais elaborados como o Expectation-Maximization (EM, hiper-elipsóides, no caso das densidades de probabilidades bases da mistura sejam Gaussianas). Também, na maioria dos casos os métodos de agrupamentos possuem elevado custo computacional. Apesar de grande parte dos algoritmos descreverem o processo como autônomo, i.e., não supervisionado, na maioria das abordagens o usuário tem grande influência sobre o resultado final do processo, através da escolha de parâmetros iniciais.

Mapas neurais constituem-se em um dos modelos mais importantes na área de pesquisa e aplicações das redes neurais, combinado aspectos de quantização vetorial com a propriedade de continuidade de funções (Bauer et al., 1999; Yin, 2008). O SOM (ou mapa auto-organizável de Kohonen), um dos principais modelos de redes neurais, sendo o mais utilizado para problemas de classificação não-supervisionada, tem sido amplamente utilizado para problemas de visualização e classificação não-supervisionada. Porém, o uso efetivo do SOM em problemas complexos (Costa, 1999b; Kohonen, 2001; Yin, 2008) demanda etapas subsequentes de processamento e análises.

O SOM define um mapeamento de um espaço p-dimensional contínuo para um conjunto finito de vetores referência, cada um representado por um neurônio da rede, dispostos na forma de um arranjo espacial regular, normalmente bi-dimensional. Todos os neurônios recebem os mesmos p sinais de entrada e o objetivo principal do treinamento é reduzir a dimensionalidade dos sinais, buscando a melhor preservação possível da topologia do espaço de entrada. O algoritmo SOM tenta ao mesmo tempo reduzir dimensionalidade (visão do espaço de saída) e efetuar agrupamento (quantização vetorial). Em um SOM treinado, espera-se que dados próximos no espaço de entrada sejam mapeados no mesmo neurônio ou em neurônios vizinhos (Kohonen, 2001; Yin, 2008b). Uma maneira de efetuar o agrupamento com SOM é rotular os neurônios com as classes dos dados mais freqüentes. Esta segmentação do espaço, de forma manual, implica além de um custo, um fato que é o uso da informação da classe no processo. Caso disponhamos da classe, originalmente, poderíamos ter utilizado algoritmos supervisionados, mais adequados, na maioria das vezes, para esta situação.

Com objetivo de obter melhor resultado da aplicação do SOM em agrupamento, métodos de pós-processamento do mapa treinado têm sido propostos, incluindo a segmentação morfológica da U-matrix, através da transformada Watershed (Costa, 1999a,b); segmentação utilizando técnicas de particionamento de grafos, por eliminação de arestas inconsistentes (Costa, 2003); segmentação do SOM por métodos de agrupamentos hierárquicos com conectividade restrita (Murtagh, 1995; Costa, 2005, Gonçalves ET al. 2008), entre outras, como aplicação do *k-means* e de métodos hierárquicos (Vesanto & Alhoniemi, 2000).

Por outro lado, índices de validação são utilizados para avaliar partições de dados produzidas por métodos de agrupamento. Além de aferir a qualidade das partições obtidas, busca-se, também, a estimativa do número de agrupamentos (Costa, 1999). Em vários casos, análises são feitas com índices após simulações variando parâmetros de algoritmos de clustering (ex. no *k-means*, variando-se o *k*).

Grande parte das abordagens de validação de agrupamentos existentes obtém melhores desempenhos quando os agrupamentos de dados são compactos e apresentam formatos hiper-esféricos (Berry e Lino, 1996; Halkidi et al., 2002). Para aplicações que apresentam agrupamentos com sobreposições e formatos arbitrários, os critérios tradicionais de validação (variância, densidade e separação) não são suficientes para realizar uma avaliação adequada (Gonçalves, 2009). Índices recentes de validação propõem mecanismos para avaliar não apenas a compacidade e a separação dos agrupamentos, mas também a geometria dos mesmos.

Halkidi e Vazirgiannis (2002) propuseram o índice CDbw (Composing Density Between and Within Clusters), que apresenta características que o torna eficiente em relação a outros índices que se baseiam apenas em critérios de compacidade e de separação dos dados. As características geométricas dos agrupamentos são representadas pelo uso de vetores representativos, isso permite que o CDbw avalie de maneira correta estruturas que não tenham apenas formas hiper-esféricas (Gonçalves, 2009, Gonçalves et al, 2008).

Este artigo apresenta uma metodologia que explora as características e propriedades do SOM para realizar visualização e agrupamento de dados utilizando segmentação de imagens, através de morfologia matemática, e índices de validação, como o CDbw simplificado. Resultados demonstram a capacidade de transformação de dados multivariados em imagens e sua segmentação, permitindo entendimento da estrutura dos mesmos.

O restante do artigo é organizado da seguinte forma: a Seção 2 descreve brevemente SOM, a matriz U e segmentação do SOM utilizando morfologia matemática. A Seção 3 descreve índices de validação modificados, como o CDbw simplificado, que são aplicados no processo de detecção de agrupamentos. A Seção 4 descreve o método proposto enquanto que a Seção 5 apresenta resultados e discussões. A Seção 6 conclui o artigo apontando também algumas possibilidades futuras.

2 Visualização e análise de agrupamentos usando SOM

O SOM é uma rede neural artificial que define um mapeamento de um espaço de entrada *p*-dimensional contínuo para um conjunto finito de protótipos (neurônios), que são dispostos em um arranjo topológico, geralmente bidimensional (Kohonen, 2001). A rede consiste essencialmente de duas camadas de neurônios. A entrada da rede corresponde a um vetor *p*-dimensional, **x**. A literatura normalmente trata o caso em que **x** é contínuo, entretanto, alguns autores exploram o caso em $\mathbf{x} \in Z^p$ ou $\mathbf{x} \in \Re^p$. Cada neurônio *i* é representado por um vetor de pesos $\mathbf{w_i} = [w_{i1}, w_{i2},..., w_{ip}]^T$, também no espaço *p*-dimensional.

Para cada padrão de entrada um neurônio é escolhido o vencedor, *c*, usando o critério de maior similaridade:

$$\|\mathbf{x} - \mathbf{w}_c\| = \min\{\|\mathbf{x} - \mathbf{w}_i\|\}, \quad (1)$$

onde ||.|| representa a distância Euclidiana. Os pesos do neurônio vencedor, juntamente com os pesos dos seus neurônios vizinhos, são ajustados de acordo com a seguinte Equação:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + h_c(t)[\mathbf{x}(t) - \mathbf{w}_i(t)], \quad (2)$$

onde *t* indica a iteração do processo de treinamento, **x**(t) é o padrão de entrada e $h_c(t)$ inclui os fatores taxa de aprendizado e núcleo de vizinhança em torno do neurônio vencedor (Kohonen, 2001), que reduzem com as épocas de treinamento.

Uma vez que a rede tenha sido treinada, o arranjo de neurônios do SOM reflete características estatísticas importantes do espaço de entrada. As principais propriedades da rede, como resumidas em (Haykin, 2001), são:

1. Aproximação do espaço de entrada: o SOM tem como objetivo básico armazenar um conjunto grande de vetores de entrada encontrando um conjunto menor de protótipos (vetores de pesos sinápticos $\mathbf{w_i}$) de modo a fornecer uma boa aproximação para o espaço de entrada original;

2. Ordenação Topológica: ao realizar o mapeamento não-linear dos vetores de entrada para o arranjo de neurônios da rede, o algoritmo do SOM tenta preservar ao máximo a topologia do espaço original;

3. Casamento de Densidade: o mapeamento efetuado pelo SOM reflete a distribuição de probabilidade dos dados no espaço de entrada original.

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). Alguns trabalhos anteriores lidando com SOM utilizaram redes pequenas com um neurônio por agrupamento (Curry et al. 2001). Essa abordagem faz com que o SOM funcione de forma similar ao *k-means*, que, usando distâncias Euclidianas, permite a descoberta de agrupamentos apenas de formas hiperesféricas.

A *U-matrix* permite visualizar as inter-distâncias entre neurônios adjacentes do SOM na forma de uma imagem (ou superfície) (Ultsch, 1993). Para um mapa com tamanho $X \times Y$ neurônios, a *U-matrix* terá tamanho $(2X-1) \times (2Y-1)$. Seja *n* um neurônio no mapa, $NN(n)$ o conjunto de vizinhos sobre o mapa, $\boldsymbol{w}(n)$ o vetor de peso associado com o neurônio *n*, então

$$U\_height(n) = \sum_{m \in NN(n)} d\big(w(n) - w(m)\big) \tag{3}$$

onde d(.) é a medida de distância que usa a métrica utilizada no algoritmo do SOM. A *U-matrix* é uma imagem dos *U-heights* sobre o topo da grade de neurônios do mapa. Valores elevados representam dissimilaridades entre neurônios e correspondem a bordas de agrupamentos, enquanto que regiões de neurônios similares apresentam valores relativamente baixos.

O uso de uma representação distribuída dos protótipos (ou vetores referência) em problemas de agrupamentos (Costa, 1999a,b) permite flexibilizar e descobrir a estrutura inerente dos dados. Para isso, há necessidade de ter um conjunto de representantes e agrupá-los (ou rotulá-los, após segmentação) de acordo com algum critério. No caso do SOM, métodos de pós-processamento do mapa treinado têm sido propostos, permitindo que agrupamentos complexos sejam identificados no mapa (Costa, 1999a,b; Vesanto & Alhoniemi, 2000).

Método de segmentação automática de SOM utilizando *U-matrix* e morfologia matemática foram apresentados em (Costa & Netto, 1999, 2001). Um ponto importante do algoritmo é a escolha dos marcadores que servirão para mudança da homotopia da imagem, durante o uso da transformada *watershed* (Meyer, 1993; Dougherty & Lotufo, 2003). Costa & Netto (1999, 2001) descreveram um método baseado em escala, que busca a maior relação contígua entre o número de regiões conectadas versus limiar a ser aplicado na *U-matrix*, Após a modificação da imagem homotopia e aplicação da *watershed*, regiões de neurônios são rotuladas de acordo com os segmentos obtidos na *U-matrix*. Variações deste algoritmo foram utilizadas em várias aplicações, por exemplo, na segmentação de imagens de satélite (Goncalves et al., 2008), agrupamento dos elementos em esquemas de integração semântica de fontes de dados heterogêneas (Zhao & Ram, 2007), determinação do número de modos de operação da base de dados multivariados SPC (Zhang & Albin, 2007) e análise de sinais de eletroencelografia (Sommer & Golz, 2001).

Técnicas de particionamento de grafos utilizando eliminação de arestas inconsistentes no SOM foram propostas em (Costa, 2003). O algoritmo apresentado em Costa (2003) foi aplicado com sucesso, com incorporação de índices de validação, por Silva et al. (2004) em análise de agrupamentos de dados geoespaciais. Variantes do método também foram aplicados em processamento de imagens multiespectrais (Gonçalves et al. 2005, 2006; Gonçalves, 2009). O uso de métodos de agrupamentos hierárquicos com conectividade restrita foi descrito em Costa (2005). O método, inspirados em Murtagh (1995), porém, utiliza o método de Ligações Simples, no lugar de métodos de centróides, o que permitiu flexibilidade na geometria dos agrupamentos descobertos. Aplicações do método hierárquico, com incorporação de melhorias, foi apresentado em Gonçalves et al. 2008; Gonçalves, 2009).

O resultado da aplicação destes algoritmos no SOM são regiões conectadas de neurônios que definem no espaço de atributos (entrada) geometrias complexas e não paramétricas. Os métodos descritos em Costa e Netto (1999, 2001), Costa (2003) e Costa (2005), 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).

Técnicas mais comuns de detecção de *clusters* usando o SOM envolvem o uso de agrupamento hierárquico aglomerativo e particionamento, ex. *k-means* (Vesanto & Alhoniemi, 2000). Além do uso destes métodos, os autores descrevem o uso de índices de validação (Davies & Bouldin, 1979) estimar o número de agrupamentos, após simulações com diferentes valores de *k*. No entanto, o processo de descoberta dos agrupamentos descrito exige muita iteração do usuário, não caracterizando, portanto, um método totalmente automático. Em simulações os métodos apresentados não foram capazes de reconhecer clusters não-esféricos, no caso do uso conjunto do SOM com o *k-means*, dada limitações da geometria dos agrupamentos obtidas, similar ao problema encontrado em (Curry et al. 2001).

Kiang (2001) também utilizou um método hierárquico aglomerativo empregando critérios de variância mínima para junção ou divisão de grupos de neurônios. O algoritmo, baseado no trabalho de Murtagh (1995), necessita recalcular os centróides sempre que dois grupos são unidos e apresenta boa aplicabilidade apenas para agrupamentos com formas hiperesféricas ou hiperelipsoidais.

Wu e Chow (2004) propuseram também um método hierárquico aglomerativo para analisar a saída do SOM utilizando um índice de validação de agrupamentos. A aplicação do método em algumas bases de dados artificiais e reais mostra que o seu desempenho é superior em relação a outras técnicas de agrupamentos convencionais.

Brugger et al. (2008) recentemente descrevem uma superfície, denominada Clusot, e métodos relacionados para tentar detectar clusters usando o SOM. A Clusot é uma superfície derivada de não apenas distâncias entre neurônios, mas também frequências de ativação dos mesmos. Máximos locais na superfície indicam um cluster. Para permitir a detecção de cluster, foram descritos

dois algoritmos, o primeiro baseado em detecção de bordas baseado em gradiente e outro baseado em inundações recursivas, considerando a superfície um relevo topográfico (similar a abordagens baseadas em *watershed*). Ambos os métodos apresentaram resultados inferiores ou equivalentes aos relatados em (Vesanto & Alhoniemi, 2000). Os métodos descritos em (Vesanto & Alhoniemi, 2000; Kiang, 2001; Brugger et al., 2008) não foram capazes de detectar automaticamente estruturas complexas, tais como os dados *chainlink* (Ultsch, 1993), que foi possível com métodos descritos em Costa & Netto (1999a, 1999b, 2001).

Uma extensão da segmentação de mapas em uma arquitetura hierárquica, orientada pelos dados, com o objetivo descobrir e representar sub-clusters, foi apresentado em Costa (1999) e Costa e Netto (2001a,b). O TS-SL-SOM (*Tree-Structured Self-Labeling SOM*) gera sub-redes para cada região rotulada de neurônios na forma de uma árvore dinâmica. Não se especifica *a priori* o número de sub-redes para uma dada rede e os parâmetros de cada sub-rede são funções dos parâmetros da rede pai e do subconjunto de dados que será usado para treiná-la. Sub-redes que não apresentam sub-partições são excluídas. Exemplos de aplicações deste método como particionamento recursivo de bases de dados também têm sido desenvolvidos (Costa et al., 2003; Costa e Netto, 2001a,b).

O método de segmentação utilizando *watershed* foi também estendido para mapas com espaço de saída com dimensão maior que 2, motivado pela manutenção da topologia, geralmente é perdida quando diminuímos a dimensionalidade via mapeamento de um espaço *p*-dimensional para um espaço de menor dimensão. O *U-array* foi definido como uma extensão da *U-matrix* e foram propostos métodos de análise baseados em morfologia matemática utilizada em redes de dimensão 1 ou 2 (Costa e Netto, 2007).

3 Índices de validação de agrupamentos

Índices de validação são utilizados para avaliar a qualidade de partições de dados produzidas por métodos de agrupamento. Busca-se, também, a estimativa do número de agrupamentos (Costa, 1999) adaptando-se corretamente aos conjuntos de dados. Em vários casos, análises são feitas com índices após simulações variando parâmetros de algoritmos de clustering (ex. no *k-means*, variando-se o *k*).

Dificilmente haverá um índice de validação que seja totalmente adequado para avaliar quaisquer tipos de agrupamentos. Dados com naturezas (características) distintas podem requerer métodos e índices de validação distintos. Cada um deles apresenta critérios de validação diferenciados e vantagens e desvantagens sob diferentes aspectos.

Na literatura podemos encontrar diferentes métodos para investigar a validação de agrupamentos de dados (Jain et al. 1999; Xu & Wunsch, 2009). A implementação da maioria dos algoritmos de validação exige alto custo computacional, principalmente quando o número de agrupamentos e a quantidade dos dados de entrada são muito grandes. Alguns índices são dependentes dos dados e/ou do número de agrupamentos. Existem índices que utilizam os centróides de subconjuntos de dados em seus algoritmos, enquanto que outros usam todos os pontos de cada subconjunto. Análises comparativas entre vários índices de validação servem de base para selecionar aqueles que apresentam melhores desempenhos na avaliação de conjuntos de dados diversos (Milligan e Cooper, 1985; Hubert e Arabie, 1985; Maulik e Bandyopadhyay, 2002; Bezdek e Pal, 1998).

Conforme descrito em Halkidi e Vazirgiannis (2002), apesar da grande variedade de índices de validação propostos na literatura, a maioria deles é computacionalmente intensiva, especialmente em aplicações que apresentam um grande volume de dados, por exemplo, análise (segmentação) de imagens médicas ou imagens de sensoriamento remoto (Gonçalves, 2009).

Apesar de termos utilizado outros índices, como o Davies-Bouldin (Davies & Bouldin, 1979), o índice que demonstrou melhor adequação (inclusive por poder trabalhar com estruturas geométricas variadas) foi o CDbw (*Composing Density Between and Within Clusters*) (Halkidi & Vazirgiannis, 2008), descrito, brevemente, a seguir.

3.1 O Índice CDbw

O CDbw (Halkidi & Vazirgiannis, 2008) apresenta características importantes em relação a outros índices de validação. O método baseia-se em dois conceitos importantes: a densidade intra-agrupamento e a separação entre os agrupamentos. A característica geométrica do agrupamento é representada pelo uso de vetores representativos, isso permite que o índice avalie de maneira correta estruturas não hiperesféricas, o que não ocorre em outros índices. Além disso, a complexidade computacional do método é $O(n)$, que é aceitável para grandes conjuntos de dados.

O cálculo do índice CDbw é feito considerando um conjunto de dados particionados em c agrupamentos. Define-se um conjunto de pontos representativos $V_i = \{v_{i1}, v_{i2}, \ldots, v_{ir}\}$ para o agrupamento i, onde r representa o número de pontos de representação do agrupamento i. A densidade intra-agrupamento é definida por:

$$Intra(c) = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{r_i}\sum_{j=1}^{r_i} density(v_{ij}), c > 1 \quad (4)$$

com $density(v_{ij}) = \sum_{l=1}^{n_i} f(x_l, v_{ij})$, onde x_l pertence ao *i*-ésimo agrupamento, v_{ij} é a *j*-ésima representação do agrupamento *i*, e $f(x_l,v_{ij})$ é dado por 1, se $||x_l - v_{ij}|| \leq stdev$, ou 0 caso contrário. *stdev* corresponde a média dos desvios padrões de cada um dos *c* agrupamentos.

A densidade inter-agrupamento é dada por

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

onde $c>1$, $close\_rep(i)$ e $close\_rep(j)$ representam o par de pontos de representação mais próximos entre o agrupamento *i* e *j*, v_{ij} é o ponto médio entre este par de pontos, $stdev(i)$ e $stdev(j)$ representam os desvios padrões dos agrupamentos *i* e *j*, e

$$density(v_{ij}) = \sum_{k=1}^{n_i+n_j} f(x_k, v_{ij}), \quad (6)$$

onde x_k pertence ao agrupamento *i* ou *j*, e $f(x_k, v_{ij})$ é dado por 1 se $||x - v|| \leq (||stedv(i) + ||stedv(j)||)$, ou 0 caso contrário.

A separação entre os agrupamentos é definida por:

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

O índice CDbw é dado então por

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

Experimentos em [18] demonstram que uma boa partição dos dados é indicada por valores altos do índice.

A modificação realizada foi seu cálculo não nos dados de entrada, mas no conjunto de neurônios, após treinamento, o que simplifica bastante o cálculo.

4 Abordagem proposta

A estratégia de análise de agrupamentos proposta neste trabalho é encontrar a melhor partição para um conjunto de dados a partir da análise de diferentes segmentações realizadas sobre a *U-matrix,* sendo uma extensão do método apresentado por Costa e Netto (1999, 2001), que emprega o algoritmo de segmentação de imagens, *watershed*, utilizando uma imagem de marcadores para regularizar o processo de segmentação. No caso atual, ao invés de definirmos apenas uma imagem de marcadores para segmentar a *U-matrix* (como apresentado em Costa e Netto (1999, 2001), várias imagens de marcadores são determinadas e, portanto, diferentes segmentações da *U-matrix* são obtidas. Cada uma dessas segmentações é associada aos neurônios do SOM, obtendo-se assim diferentes partições do mapa de neurônios que definem diferentes configurações de agrupamentos dos dados. Para cada uma das diferentes configurações de agrupamentos são aplicados os índices de validação CDbw, que servem como critério de decisão para a escolha da melhor partição dos dados.

Considerando a U-matrix de um SOM treinado dada pela imagem U com 256 níveis de cinza, os seguintes passos são efetuados para se obter o conjunto de imagens de marcadores para *U*:

1. Filtragem: obter uma imagem U_1 suavizando a imagem *U* através da remoção de pequenas depressões com área inferior a 3 pixels;

2. Para $k = 1,...,f_{max}$, onde f_{max} é o nível de cinza máximo na imagem U_1, criar as imagens binárias U_2^k correspondendo a conversões de U_1 usando k como valor de limiar;

3. Obter o número de regiões conectadas de U_2^k, para cada valor de k, N_{rc}^k;

4. Construir o gráfico $k \times N_{rc}^k$, e através dele obter o conjunto S_k composto por todos os valores de k que correspondem a inícios de seqüências contíguas e constantes de números de regiões conectadas com tamanhos superiores a 3;

5. Obter o conjunto de todas as imagens de marcadores, $S_m = \{ U_2^{k_1}, U_2^{k_2}, ..., U_2^{k_n} \}$, onde $k_1, k_2, ..., k_n$ são os elementos de S_k obtidos no passo 4.

Embora o procedimento usado para encontrar os marcadores da *U-matrix* seja semelhante ao apresentado em Costa e Netto (1999, 2001), na abordagem atual os passos 4 e 5 determinam um conjunto de imagens de marcadores da *U-matrix* considerando todas as seqüências contíguas e constantes de regiões conectadas. Em todos os experimentos foram considerados apenas as seqüências com tamanhos maiores que 3.

Sendo assim, a estratégia geral de particionamento de um conjunto de dados, a partir de um SOM treinado com sucesso, pode ser resumida como segue:

1. Obtenha a *U-matrix* a partir do SOM treinado;
2. Encontre o conjunto, $S_m = \{ U_2^{k_1}, U_2^{k_2}, ..., U_2^{k_n} \}$, de imagens de marcadores para a *U-matrix*;
3. Para cada imagem de marcadores, $U_2^{k_i}$, $i=1,2,..n$, faça:
    a) Aplicar o algoritmo *watershed* sobre a *U-matrix*;
    b) Rotular as regiões conectadas da imagem segmentada no passo 3.a;
    c) Realizar a cópia dos rótulos obtidos no passo 3.b para os neurônios do SOM associados a cada pixel da *U-matrix*;
    d) Rotular o conjunto de dados de entrada utilizando o SOM rotulado no passo 3.c;
    e) Aplicar um índice de validação sobre o conjunto de dados rotulados obtido no passo 3.d.
4. Escolher a partição de dados que apresenta o melhor valor para o índice de validação utilizado.

Diferentemente da proposta apresentada em (Halkidi & Vazirgiannis, 2008), que encontra os vetores de referência dos agrupamentos de forma iterativa a partir do conjunto de dados, em nossa abordagem os próprios vetores de pesos dos neurônios do SOM são utilizados como os vetores de referência dos seus respectivos agrupamentos, simplificando assim o cálculo do índice.

5 Resultados

Esta Seção apresenta resultados da abordagem proposta na análise de agrupamentos aplicada no conjunto de dados Wine (Asuncion & Newman, 2007). Os dados são resultados de uma análise química de vinhos produzidos em uma mesma região da Itália, mas derivados de três formas de cultivo diferentes. São 178 amostras divididas em 3 classes: 59 pertencentes a classe 1, 71 pertencentes a classe 2 e 48 pertencentes a classe 3. Cada amostra possui 13 atributos. A Figura 1 mostra a projeção dos dados amostrais rotulados utilizando os dois componentes principais.

Experimentos foram feitos com SOM com tamanhos variados, porém com as seguintes configurações: topologia retangular, inicialização dos pesos linear, vizinhança tipo Gaussiana, algoritmo em lote, 500 épocas de treinamento e raio de vizinhança inicial 80% do tamanho do mapa e vizinhança final 1. As Figuras 2 e 3 ilustram o histograma de ativação e *U-matrix* para um mapa de tamanho 10x10. O gráfico do número de regiões conectadas versus o limiar (k) da *U-matrix* é apresentado na Figura 4. A Tabela 1 mostra todos os valores de k que correspondem a inícios de seqüências contíguas, juntamente com os números de agrupamentos correspondentes.

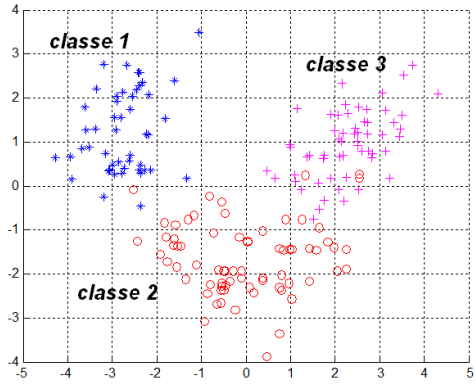

Figura 1: Projeção do conjunto de dados Wine utilizando os dois primeiros componentes prinicipais.

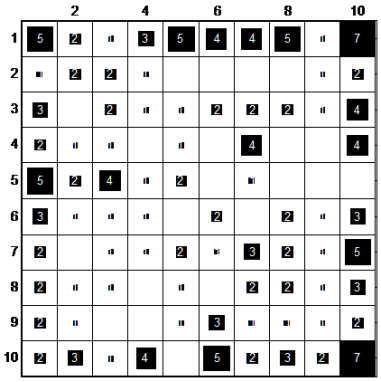

Figura 2: histograma de ativação, mapa de tamanho 10x10.

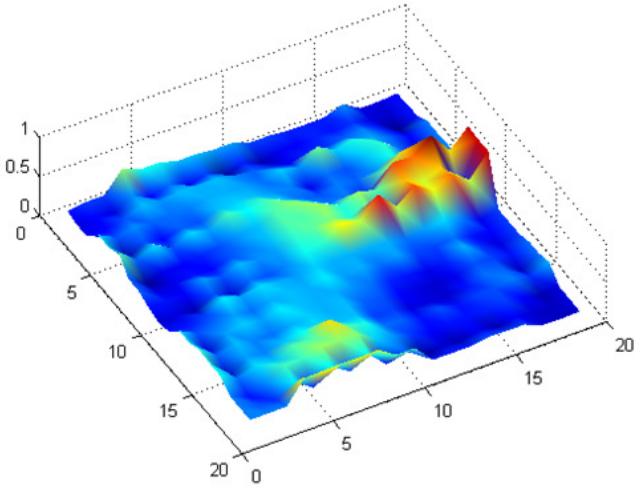

Figura 3: *U-matrix*, mapa de tamanho 10x10.

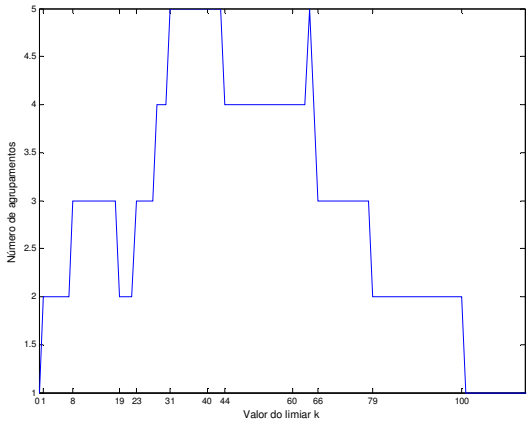

Figura 4: Número de regiões conectadas versus o limiar (*k*) da *U-matrix*.

Para cada um dos valores de *k* da Tabela 1 foram determinadas então as imagens de marcadores da *U-matrix* e aplicada a estratégia de particionamento dos dados. O valor de *k* que apresenta maior estabilidade possui valor igual a 79 e corresponde a 02 agrupamentos. Este é o valor de *k* utilizado pelo método proposto por Costa-Netto (Costa & Netto, 1999; 2001) para realizar o particionamento dos dados.

A melhor partição do conjunto de dados foi obtida para o índice CDbw com valor 34.142 (maior valor dentre os demais). O valor de *k* correspondente é igual a 66, que de acordo com a Tabela 1, equivale a 3 agrupamentos. As Figuras 5 e 6 ilustram o processo de particionamento dos dados considerando o valor de limiar *k* igual a 66, que gera a melhor partição do conjunto de dados de acordo com o índice CDbw. A Figura 7 ilustra o SOM rotulado enquanto que a Figura 8 ilustra o particionamento obtido dos dados com a segmentação pelo método proposto.

Tabela 1: valores do limiar *k* que correspondem a inícios de seqüências estáveis, números de agrupamentos e índices CDbw correspondentes.

Valores de *k* (inícios de seqüências estáveis)	Número de agrupamentos	Índice CDbw
1	2	14.31
8	3	14.55
19	2	13.38
23	3	30.97
31	5	5.53
44	4	14.96
66	3	34.14
79	2	19.50

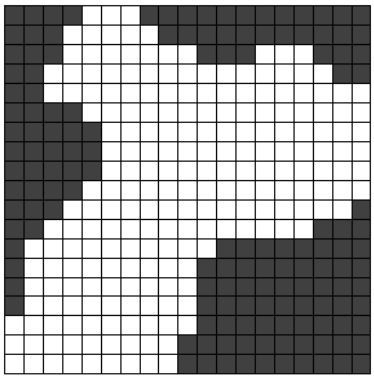

Figura 5: Imagem de marcadores para a *U-matrix* usando como limiar k = 66.

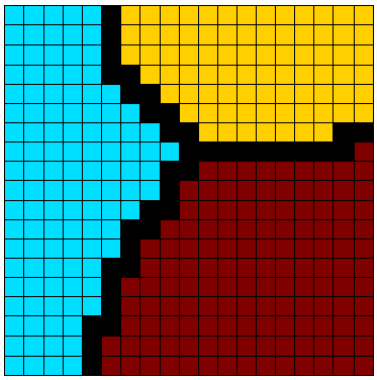

Figura 6: Linhas e segmentos da *U-matrix* obtidas usando *watersheds* e os marcadores da Figura 5.

Observando as Figuras 6 a 8, verifica-se que o conjunto de dados foi particionado em três classes, para um k = 66, o qual corresponde ao maior valor do índice CDbw. A acurácia entre os rótulos das classes descobertas pelo método de *clustering* e das classes reais (não utilizadas durante o processamento) pode ser obtida, de forma a gerar um indicador que reflita a capacidade do método de encontrar os grupos naturais dos dados.

No caso apresentado nas Figuras 6 a 8, o método com CDbw obteve índice de acerto de 96.63%, um pouco melhor que quando utilizamos o índice Davies-Bouldin, que ficou em 92%. Os erros de classificação são devidos a problemas com os dados, como sobreposição dos agrupamentos, devido à má escolha dos atributos para representar as classes ou erros de medição.

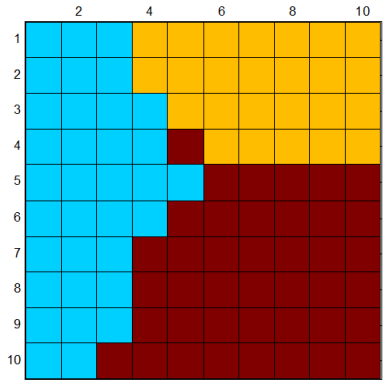

Figura 7: SOM rotulado.

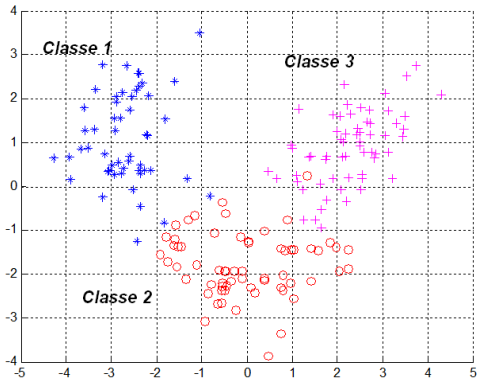

Figura 8: Particionamento obtido dos dados com a segmentação pelo método proposto, limiar k =66.

A Tabela 2 mostra valores de acurácia, usando o método proposto, para diferentes tamanhos do SOM. Nota-se que a acurácia decai para mapas muito pequenos (ex. 4x4) e para mapas grandes (ex. 30x30). Ressalta-se que o parâmetro filtro utilizado no método *area_open* e *area_close* (Dougherty & Lotufo, 2003) para pré-processar a *U-matrix* deve aumentar de acordo com o tamanho do mapa, tendo sido utilizado o valor médio de uma das dimensões do mapa.

Tabela 2: Acurácia para diferentes tamanhos do SOM, usando o método proposto.

Tamanho do mapa	Acurácia
4 x 4	76.97%
5 x 5	96.07%
8 x 8	92.13%
9 x 9	93.82%
10 x 10	96.63%
15 x 15	94.38%
20 x 20	92.70%
30 x 30	91.57%

Para um mapa de tamanho 10x10, simulando com o método descrito em Vesanto & Alhoniemi (2000), com uso de *k-means*, em 10 simulações para *k* variando de 2 a 10, e quando *k* foi selecionado com valor 3, a acurácia obtida foi 95.51%, no melhor caso.

6. Conclusões

Apresentou-se uma metodologia que explora as características e propriedades do SOM para realizar visualização e agrupamento de dados utilizando segmentação de imagens, com uso da transformada *watershed* e o índices de validação CDbw simplificado. Resultados demonstraram a capacidade de transformação de dados multivariados em imagens e sua segmentação automática, permitindo entendimento da estrutura dos mesmos.

Dado o volume e complexidade das bases de dados nas mais diversas aplicações atuais, esta ferramenta pode ser de grande utilidade para processar grandes massas de dados não rotuladas. Trabalhos futuros podem explorar diferentes índices de validação e seu uso com diferentes imagens (superfícies) como a descrita em Brugger et al. (2008).

7. Agradecimento

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

8. Referências

Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Academic Press.

Arabie, P., Hubert, L. J., De Soete, G. (1999). *Clustering and Classification*, World Scientific.

Asuncion, A. & 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]

Bauer, H.-U., Herrmann, M. and Villmann, T. (1999). Neural maps and topographic vector quantization. *Neural Networks*, Volume 12, Issues 4-5, pp. 659-676.

Berry, M.J.A. e Linoff, G. (1996). *Data mining techniques for marketing, sales and customer support*, 2a. ed. (John Willey and Sons, Inc).

Bezdek, J.C. e Pal, N.R., (1998). Some new indexes of cluster validity. *IEEE Tran. on Systems, Man and Cybernetics*, Vol. 28, pp. 301-315.

Brugger, D., Bogdan, M. & 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. & 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. & 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. (1999*). Classificação automática e análise de dados por redes neurais auto-organizáveis.* Tese de Doutorado, Departamento de Engenharia de Computação e Automação Industrial, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas.

Costa, J.A.F. (2005). Segmentação do SOM por Métodos de Agrupamentos Hierárquicos com Conectividade Restrita. VII *Congresso Brasileiro de Redes Neurais* (CBRN), Natal, RN, outubro de 2005.

Costa, J.A.F. and Netto, M.L.A. (2001), A new tree-structured self-organizing map for data analysis. In: *Proc. of the Intl. Joint Conf. on Neural Networks* (IEEE), Washington, DC, July 2001, pp. 1931-1936.

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

Costa, J.A.F., e Netto, M.L.A. (2003) Segmentação do SOM Baseada em Particionamento de Grafos. *Anais do VI Congresso Brasileiro de Redes Neurais*, São Paulo, Junho de 2003, pp. 451-456.

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.

Curry, B., Davies, F., Phillips, P., Evans, M. & Mouthino, L. (2001). The Kohonen self-organizing map: an application to the study of strategic groups in the UK hotel industry. *Expert Systems*, 18 (1). pp. 19-31.

Davies, D.L. & Bouldin, D.W. (1979). A cluster separation measure. *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 1, 4, pp. 224-227.

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

Everitt, B. S., Landau, S., Leese, M. (2001). *Cluster Analysis*, Hodder Arnold Publication.

Gan, G., Ma, C., WU, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability, SIAM.

Goncalves, M., Netto, M., Zullo, J. & Costa, J.A.F. (2008). A new method for unsupervised classification of remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods. *In: Intl. Journal of Remote Sensing*, Vol. 29, Issue 11, June 2008, pp. 3171 – 3207.

Gonçalves, M.L. (2009). *Métodos de Classificação Não-supervisionada de Imagens de Sensoriamento Remoto Usando Mapas Auto-organizáveis de Kohonen*. Tese (Doutorado em Engenharia Elétrica), Universidade Estadual de Campinas, SP.

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

Halkidi, M. e Vazirgiannis, M., (2002). Clustering validity assessment using multi representatives. *In: Proceedings of the SETN Conference*, Thessaloniki, Grécia.

Hartigan, J. A. (1975). *Clustering Algorithms*, John Wiley & Sons.

Haykin, S. (2001). *Redes neurais: princípios e práticas*. Porto Alegre: Bookmann, 2 ed.

Hubert, L.J. e Arabie, P., (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.

Jain, A. K., Dubes, R. C. (1988). *Algorithms for Clustering Data*, Prentice Hall.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323

Kaufman, L., Rousseeuw, P. J. (1990). *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley.

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis*, v. 38, pp. 161-180.

Kogan, J. (2006). *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press.

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

Maulik, U. e Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24, 1650-1654.

Meyer, F. (1993). Gradients and watershed lines. In: Serra, J. & Salembier, P. (Eds.), *Proc. Workshop on Mathematical Morphology and its Applications to Signal Processing*, Barcelona, pp. 70-75.

Milligan, G.W. e Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, pp. 159-179.

Murtagh, F. (1995). Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters*, vol. 16, pp. 399-408.

Sahmer, K., Vigneau, E. and Qannari, El M. (2006). A cluster approach to analyze preference data: Choice of the number of clusters. *Food Quality and Preference*, V. 17, [3-4], PP. 257-265.

Sommer, D. & Golz, M. (2001). Clustering of EEG-Segments Using Hierarchical Agglomerative Methods and Self-Organizing Maps. *Lecture Notes in Computer Sciences*, Springer, vol. 2130, pp. 642-649.

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..

Vesanto, J. & 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. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Ed., Morgan Kaufmann.

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

Xu, R. and Wunsch II, D. (2005). *Survey of Clustering Algorithms. IEEE Transactions on Neural Networks*, Vol.16, Iss.3, pp. 645-678.

Yin. H. (2008). Self-organising maps: Background, theories, extensions and applications. In: *Computational Intelligence: A Compendium*, Springer, pp. 715-762.

Younis, O. and Fahmy, S. (2004). "HEED: A hybrid, energy-efficient, distributed clustering approach for ad-hoc sensor networks", *IEEE Transactions on Mobile Computing*, 3(4):366-379, Oct-Dec.

Zhang, H. & Albin, S. (2007). Determining the number of operational modes in baseline multivariate SPC data. *IIE Transactions*, 39 (12), pp. 1103-1110.

Zhao, H. & Ram, S. (2007). Combining schema and instance information for integrating heterogeneous data sources. *Data and Knowledge Engineering*, 61, pp. 281-303.