Clodis Boscarioli

Análise de Agrupamentos baseada na
Topologia dos Dados e em
Mapas Auto-organizáveis

Tese apresentada à Escola Politécnica
da Universidade de São Paulo para
obtenção do Título de Doutor em
Engenharia Elétrica.

São Paulo
2008

Clodis Boscarioli

Análise de Agrupamentos baseada na
Topologia dos Dados e em
Mapas Auto-organizáveis

Tese apresentada à Escola Politécnica
da Universidade de São Paulo para
obtenção do Título de Doutor em
Engenharia Elétrica.

Área de concentração:
Sistemas Eletrônicos

Orientador:
Prof. Dr. Emilio Del Moral Hernandez

São Paulo
2008

FICHA CATALOGRÁFICA

Boscarioli, Clodis

Análise de Agrupamentos baseada na Topologia dos Dados e em Mapas Auto-organizáveis. São Paulo, 2008. 118 p.

Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Descoberta de conhecimento 2.Mineração de dados 3.Análise de agrupamentos 4.Redes neurais I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos. II.t.

À Deus e aos espíritos superiores que me acompanham e orientam nesta vida, e
à minha família, hóspedes eternos no meu coração!

Agradecimentos

Aos meus pais e primeiros mestres, Clélia e Laércio, que me apoiaram em todos os caminhos que escolhi percorrer e que, mesmo sem muitas vezes compreenderem meus propósitos, sempre aceitaram minhas escolhas.

Aos meus irmãos Cássia e Cleiton e ao meu cunhado Antonio, por demonstrarem, de maneira admirável, como é possível manter unida uma família, estando eu sempre geograficamente distante.

Aos meus sobrinhos Camila e Flávio e ao meu afilhado Luka, crianças lindas com quem tanto gostaria de conviver mais, acompanhando o crescimento.

Além da família que amo demais, são tantas as pessoas a agradecer nessa fase de doutoramento, que chega a ser difícil. Sem meus amigos, os de todos os tempos e os conquistados a partir de 2003 - ano em que cheguei em São Paulo, na USP, sem conhecer nada e ninguém, acho que essa jornada não teria sido possível.

Aos ex-professores Marcos Sfair Sunye e Laura Sanchés García (UFPR) e Maria Salete Marcon Gomes Vaz (UEPG), por terem fornecido boas cartas de recomendação, que certamente, foram fundamentais no processo de seleção.

Ao meu orientador, Prof. Emilio Del Moral Hernandez pela confiança depositada, pelo apoio e exemplo de profissional e ser humano. Sempre me lembrarei de você chegando com Virgínia no HU-USP, sábado à noite, após saber que eu seria submetido à cirurgia de fratura de tornozelo. Obrigado pelo companheirismo e preocupação. É um prazer fazer parte do ICONE (Grupo de Inteligência Computacional, Modelagem e Neurocomputação), sob sua orientação.

À Sarajane Marques Peres, amiga de todas as horas e tantas afinidades. Muito importantes foram as nossas longas e sempre proveitosas conversas via Internet durante e sobre nossos doutorados. Obrigado pela sempre pronta disposição em colaborar e por ter-se proposto a ler partes do que veio a tornar-se esse documento.

À minha amiga Lucia Fabíola da Silva. Veterana nos tempos da graduação. Colega. Amiga de trabalho. Amiga. Ex-aluna em pós-graduação *lato sensu*. Amiga. Parceira de viagem. Amiga. Várias foram as formas que interagimos desde que nos conhecemos, mas a nossa amizade sincera e sempre presente só aumentaram.

Aos amigos e compadres Lucio Mauro e Karina Machado e aos amigos Anne Gigliola Silva, Ezequiel Gueiber e Benifrancis Matias. Além de amigos fiéis, vocês são referências constantes pelas quais sempre quero retornar a Ponta Grossa para reencontros.

À Patrícia Junqueira Grandino, grande amiga que tanto admiro também pela trajetória profissional. Amiga de ir ao cinema, sair para dançar e principalmente, de poder contar. Sem palavras de agradecimento pela acolhida em sua família desde o momento em que saí do hospital até voltar a andar. Além de sua amizade, a convivência com Fernanda e Pedro e D. Amélia foi-me um presente.

Agradeço também à Nilza, que de bom grado e carinhosamente incluiu minha recuperação em suas atividades diárias de secretária do lar, ajudando-me “nos tempos do andador” com os banhos de contraste, administração dos remédios etc. Certamente esse cuidado acelerou minha recuperação.

Ao Alexandre César Tavares Vidal, meu irmão de coração e um dos amigos mais cultos e de humor impagável que tenho. Obrigado pelos “tempos de república”, pelo carinho fraterno e pelo seu exemplo de perseverança.

À Priscila Braga Calíope, amiga de laboratório que gosto muito. Obrigado por ter-me apresentado a Emilio e ao ICONE, e pela convivência sempre agradável no laboratório.

Ao “pai dos gêmeos”, meu amigo Marcel Jaques Simonette, pelos nossos “cafés com bobagem”, almoços de papos sérios sobre nossas preocupações, anseios e impressões sobre cursar pós-graduação, que em muito me ajudaram em momentos de desânimo ou cansaço.

Ao grande amigo de laboratório e de “tomar uma cerveja no final do expediente”, Leandro Augusto da Silva, em muito agradeço: pela amizade, pela convivência no Grupo de Pesquisa, por ter revisado parte dessa tese, pela ajuda

com MatLab e pelas tantas horas que estudamos juntos e/ou, discutimos nossas pesquisas de doutorado.

À minha amiga Claudia de Oliveira Melo, por estar sempre presente, mesmo que nos últimos tempos à distância. Nossa empatia foi imediata. Altos papos, bons momentos e apoio mútuo. Além de amigos, temos em comum o fato cabuloso de ambos termos quebrado o pé durante a pós-graduação. Valeu as dicas variadas sobre cuidados na fase de recuperação. Obrigado também pela revisão de alguns capítulos dessa tese.

Ao Marcelo Facio Palin, um grande amigo e um exemplo de bondade, por quem tenho muito carinho e respeito. Agradeço a amizade, as nossas conversas sobre docência e sobre o futuro, e pelas dicas sobre Latex.

Ao casal Diana Maia e Gustavo Esteves, amigos pelos quais nutro um carinho muito grande. Eles foram também os responsáveis pela minha introdução à capoeira.

À Déborah Schneider, Neidinha, Tutty, Sara A. de Medeira, Roberta R. Parcianello, Ivone Rojo e Luiz Antonio Rodrigues, bons amigos de Cascavel. Retornar tornou-se mais fácil pelo fato de vocês estarem aí. Déborah e D. Ivone, obrigado também por terem me acolhido antes de novamente estabelecer residência em Cascavel.

Aos meus colegas de trabalho da Universidade Estadual do Oeste do Paraná por muito me incentivarem e apoiarem. Em especial, a Jorge Bidarra e Claudia Brandelero Rizzi pela “cobrança” pelo término desse trabalho, sempre evidenciando sua importância ao nosso Grupo de Pesquisa, o GIA (Grupo de Inteligência Aplicada), que caminha para sua consolidação. E à própria Instituição, por ter-me liberado por três anos das atividades docentes para dedicar-me ao Doutorado.

Aos professores com os quais tive contato durante esse período em disciplinas e/ou seminários, o meu muito obrigado. Aprendi muito com vocês e alguns certamente me servirão de exemplo na trajetória acadêmica. Destaco o Prof. Dr. João Eduardo Ferreira, o jef, um entusiasta nos Sistemas de Banco de Dados, com quem muito aprendi nas disciplinas cursadas, como bolsista em Projeto PAE ou nos seminários de seu Grupo de Pesquisa em Bancos de Dados no IME-USP.

Aos funcionários da EP, IME e EACH, pelo carinho e profissionalismo com que sempre me trataram. Destaco entre eles, Claudia Ferreira de Souza Leite, Solange Paixão, Célia Maria de Lima, Marcelo Vantini, Adriano de J. Alves Ferreira, Lucileide Tadei e Maximiano Gonçalves Oliveira. Faço um agradecimento especial à Valquiria Alaminos, que com o tempo tornou-se uma amiga muito querida.

Durante esse período de doutoramento também conheci muitas outras pessoas legais, inteligentes e interessantes, que também cursavam pós-graduação, que fizeram-se presentes nesse período. Entre elas, Arlindo F. Conceição, Daniela Finco, Elier B. Cristo, Ana Beatriz V. Graciano, Jihan Zoghbi, Mehran Misaghi, Giulian D. Luz, Giuliano Mega, Ney Barros Luz, Raphael Y. de Camargo, Silvia Calbo, Silvia Cristina Pinto, Rosianni Cruz, Renato da S. Ramalho, Humberto Sandamann, Celina M. Takemura, Eyder Martinez Montoya, Rodrigo Fresneda, Ricardo L. Andrade, José Ribamar Braga Jr., Karina Valdívia, Kelly Rosa Braghetto, Eduardo Takeo Ueda, Ana K. Simões e Rosimeiri Toyota. A convivência e o compartilhar diferentes experiências foram-me muito gratificantes.

Aos meus amigos das comunidades “Trilheiros & Bêbados”, “Equipe Pé na Lama” e seus agregados, pelas trilhas e contato consciente e saudável com a natureza que ajudaram-me a manter a serenidade. Em especial a Claudio Sekiguti, que literalmente salvou-me a vida, quando em uma queda fiquei pendurado em uma cachoeira. Também aos amigos que saíram das atividades dos grupos para uma grande amizade: Roberval Mela Jr., Kariny Ribeiro, Luiza M. Rocha, Adriana Luchini, Ana Paula Pazini, Suzete Pazini, Andressa V. Robes, Adriana Martins e Gabriel Cordeiro Lima. A preocupação e o cuidado de vocês quando da cirurgia e pós-operatório são inesquecíveis.

Gostaria também de agradecer a todas as pessoas que de alguma forma passaram pela minha vida nesse período. Como não acredito em acasos, certamente esses “encontros” colaboraram para tornar-me a pessoa que hoje sou.

Sinceramente, estou muito feliz ao terminar essa seção de agradecimentos, por ter me dado conta do tanto que eu tenho para agradecer a Deus!

Resumo

Cada vez mais, na conjuntura das grandes tomadas de decisões, a análise de dados massivamente armazenados se torna uma necessidade das mais variadas áreas de conhecimento. A análise de dados envolve a realização de diferentes tarefas, que podem ser realizadas por diferentes técnicas e estratégias como análise de agrupamento de dados. Esta pesquisa enfatiza a realização da tarefa de análise de agrupamento de dados (*Data Clustering*) usando SOM (*Self-Organizing Maps*) como principal artefato. SOM é uma rede neural artificial baseada em aprendizado competitivo e não-supervisionado, o que significa que o treinamento é inteiramente guiado pelos dados e que os neurônios do mapa competem entre si. Essa rede neural possui a habilidade de formar mapeamentos que quantizam os dados, preservando a sua topologia. Este trabalho introduz uma nova metodologia de análise de agrupamentos a partir de SOM, que considera o mapa topológico gerado por ele e a topologia dos dados no processo de agrupamento. Uma análise experimental e comparativa é apresentada, evidenciando a potencialidade da proposta, destacando, por fim, as principais contribuições do trabalho.

Abstract

More than ever, in environment of large decision-making, the analysis of data stored massively becomes a real need in almost all knowledge areas. The data analyzing process covers the performing of different tasks that can be executed for different techniques and strategies like the data clustering analysis. This research is focused on the analysis task of data groups, called Data Clustering using Self Organizing Maps (SOM) as the principal artifact. SOM is an artificial neural network based on competitive and unsupervised learning, what means that its training is entirely driven by the data, such the neurons of the map compete themselves for doing it. This neural network has the ability to build the mapping task that quantifies the source data, but preserving the topology. This work introduces a new clustering analysis methodology based on SOM, considering the topological map produced by it and also the topology of the data obtained in the clustering process. The experimental and comparative analysis are also presented to demonstrate the potential of the proposal, highlighting at the end the main contributions of the work.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

Convenções e Lista de Símbolos

1	Introdução	1
1.1	Motivação	2
1.2	Objetivos	3
1.3	Contribuições Esperadas	4
1.4	Organização do Trabalho	4
2	Análise de Agrupamento de Dados	6
2.1	Conceitos Gerais	6
2.2	Medidas de Proximidade	12
2.2.1	Proximidade entre dois Objetos	12
2.2.2	Proximidade entre dois Grupos	15
2.2.3	Proximidade intra Grupos	16
2.3	Categorias de Algoritmos de Agrupamento	17
2.4	Validação de Agrupamentos	25
2.5	Considerações Finais	27

3	Mapas Auto-Organizáveis	29
3.1	Redes Neurais Artificiais e Princípios de Auto-organização	29
3.2	Conceitos Básicos do SOM	32
3.2.1	O Processo de Treinamento	38
3.2.2	Questões sobre Parametrização do Mapa	41
3.3	Propriedades de um SOM	44
3.4	Medidas Avaliativas para um SOM	45
3.4.1	Quantização Vetorial	45
3.4.2	Topologia	46
3.5	Considerações Finais	50
4	Análise Exploratória via Mapa Auto-Organizável	52
4.1	Conceitos Gerais	52
4.2	Visualização de um Mapa Auto-Organizável	56
4.3	Agrupamento de dados por Mapa Auto-Organizável	60
4.3.1	SL-SOM <i>Self-Labeled SOM</i>	60
4.3.2	Metodologia de Vesanto & Alhoniemi	62
4.3.3	Agrupamento baseado em Particionamento de Grafos	63
4.3.4	Agrupamento utilizando um Índice de Validação Local	65
4.3.5	Agrupamento considerando Topologia	66
4.4	Metodologia Proposta	68
4.5	Considerações Finais	72
5	Avaliação Experimental	73
5.1	Experimento Introdutório ilustrando a Proposta	74
5.2	Experimentos Comparativos sobre Bases de Dados Rotuladas	82

5.2.1	Iris Plants	82
5.2.2	Wisconsin Breast Cancer	84
5.2.3	Dermatology Database	87
5.3	Experimentos de Análise Exploratória - Aplicações em Novas Bases de Dados Reais	89
5.3.1	Base de dados de Argila do Xingu	90
5.3.2	Base de dados de Cerâmica Marajoara	92
5.4	Discussão Geral sobre os Resultados	96
5.5	Considerações Finais	98
6	Conclusões	100
6.1	Principais Considerações	100
6.2	Principais Contribuições	103
6.3	Trabalhos Futuros	104
6.4	Publicações	106
	Anexo A – Algoritmos de Validação	107
A.1	Índice Davies-Bouldin	107
A.2	Índice CDbw	108
	Referências Bibliográficas	110

Lista de Figuras

2.1	Etapas de um processo de análise de agrupamentos.	8
2.2	Exemplos de geometrias de grupos. (a) Grupos compactos. (b) Grupos alongados. (c) Grupo esférico e elipsoidal.	10
2.3	Diferentes grupos formados sobre um mesmo conjunto de dados. .	11
2.4	Definição visual de proximidade entre grupos.	16
2.5	Exemplo de um agrupamento hierárquico aglomerativo.	19
3.1	Tipos de estrutura de um mapa auto-organizável.	33
3.2	Arquitetura de um SOM 2D.	35
3.3	Ilustração da Vizinhança Gaussiana no Mapa.	37
3.4	Atualização do neurônio vencedor e de seus vizinhos.	38
3.5	Exemplo de um mapa sendo treinado para dados e protótipos bidimensionais. O mapa está inicialmente ajustado em uma região não superposta à dos dados, para evidenciar a movimentação dos seus protótipos para se ajustar aos dados, interativamente.	41
4.1	Análise de Dados a partir do SOM.	53
4.2	Mapa topológico gerado por SOM. As coordenadas x_1 e x_2 representam as dimensões dos vetores de dados de entrada. Cada ponto representa um protótipo no mapa.	54
4.3	Distribuição dos dados de entrada.	55
4.4	Mapa SOM rotulado, onde letras indicam rótulos e os números entre parênteses a contagem dos vetores de dados representados por cada protótipo.	56

4.5	Exemplo de construção da Matriz-U para uma estrutura de vizinhança retangular, onde se vê as três distâncias calculadas.	57
4.6	Esquema de preenchimento e plotagem de uma Matriz-U.	58
4.7	Exemplos de Matriz-U: (a) Matriz bidimensional em tons de cinza; (b) Matriz visualizada com representação por superfície topológica.	59
4.8	Diagrama de blocos para a metodologia proposta.	69
4.9	Ilustração da construção de uma matriz de ligações.	70
5.1	Base de dados artificiais. Em (a), a configuração dos dados de entrada no espaço bidimensional é apresentada. Em (b), é dada a sobreposição dos neurônios da rede SOM aos dados, rotulados os dados pelo número de seu neurônio representante no mapa.	75
5.2	Dendrograma inicial dos 30 neurônios vencedores.	76
5.3	Avaliação do primeiro nível do dendrograma. Em (a), é evidenciado o ramo esquerdo. Em (b), segue a matriz de ligações resultante.	77
5.4	Avaliação do primeiro nível do dendrograma. Em (a), é evidenciado o ramo direito. Em (b), a matriz de ligações resultante é dada.	77
5.5	Dendrograma final dos neurônios vencedores.	78
5.6	Matrizes de ligação para o segundo nível do dendrograma. Em (a), a matriz condizente ao ramo de cor azul. Em (b), a matriz de ligações do ramo de cor vermelha. Em (c), a matriz do ramo de cor verde.	79
5.7	Mapa SOM final rotulado.	80
5.8	Agrupamento encontrado pela metodologia proposta.	81
5.9	Agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.	81
5.10	Agrupamentos encontrados para a base de dados <i>Iris</i> com mapa 4×4 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.	83

5.11	Agrupamentos encontrados para a base de dados <i>Iris</i> com mapa 5×5 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.	84
5.12	Agrupamentos encontrados para a base de dados <i>Wisconsin Breast Cancer</i> com mapa 7×7 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi. B - Benigno; M - Maligno.	86
5.13	Agrupamentos encontrados para a base de dados <i>Wisconsin Breast Cancer</i> com mapa 9×9 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi. B - Benigno; M - Maligno.	86
5.14	Agrupamentos encontrados para a base de dados <i>Dermatology</i> com mapa 9×9 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.	88
5.15	Mapas rotulados para os dados do Xingu. Em (a), a visualização dos dados dos depósitos de argila <i>A</i> e <i>B</i> no mapa. Em (b), a classificação dos dados de cerâmica junto aos depósitos.	91
5.16	Agrupamentos encontrados para a base de dados Marajoara pela metodologia proposta.	94
5.17	Agrupamentos encontrados para a base de dados Marajoara pela metodologia de Vesanto & Alhoniemi.	94
5.18	Gráfico da análise discriminante.	95

Lista de Tabelas

2.1	Exemplos de algoritmos de agrupamento, por categorias.	22
5.1	Distribuição da base de dados artificial em classes.	74
5.2	Distribuição da base de dados <i>Iris</i> em classes.	82
5.3	Medidas de avaliação para os mapas da base de dados <i>Iris</i>	83
5.4	Distribuição da base de dados <i>Wisconsin Breast Cancer</i> em classes.	85
5.5	Medidas de avaliação para os mapas da base de dados <i>Wisconsin Breast Cancer</i>	87
5.6	Distribuição da base de dados <i>Dermatology</i> em classes.	88
5.7	Medidas de avaliação para os mapas da base de dados <i>Dermatology</i> .	88

Lista de Abreviaturas

ART *Adaptative Resonance Theory*

BMU *Best Matching Unit*

CDbw *Composing Density Between and With Clusters*

DB *Índice Davies-Bouldin*

INAA *Instrumental Neutron Activation Analysis*

IPEN Instituto de Pesquisas Energéticas e Nucleares

MAE Museu de Arqueologia e Etnologia

ML Matriz de Ligações

PCA *Principal Components Analysis*

ppb Partes por bilhão

ppm Partes por milhão

RNAs Redes Neurais Artificiais

SL-SOM *Self-Labeled SOM*

SOM *Self-Organizing Map*

SVMs *Support Vector Machines*

WTA *Winner-Takes-All*

Convenções e Lista de Símbolos

Na notação das fórmulas, há as seguintes convenções:

- Letras maiúsculas representam conjuntos, tais como X e Y ;
- Letras minúsculas indicam elementos de conjunto, como x, y ;
- Subscrito diferencia um elemento de um conjunto, por exemplo, x_i, y_i e w_i ;

Os seguintes símbolos são utilizados nesta tese:

E_q	Erro de Quantização
E_t	Erro Topológico
P_t	Produto Topológico
n	Cardinalidade (Número de elementos) de um conjunto de dados
p	Número de dimensões (variáveis) de um vetor de dados
d	Métrica de dissimilaridade
d_i	Dimensão de um mapa de saída SOM
s	Medida de similaridade
$\Re$	Conjunto dos números reais
D_k	Representação de um agrupamento de dados
n_k	Número de elementos de um agrupamento de dados C_k
k	Número de grupos a formar
c_k	Centro de um grupo
u	Designação de um neurônio
t	Variável que indica tempo
$\alpha(t)$	Taxa de aprendizado
$\sigma(t)$	Raio da vizinhança no tempo t
$\rho(t)$	Função decrescente no tempo
h_{bmu}	Função de vizinhança do neurônio BMU

continua...

N_c	Neurônios que sofrerão atualização
i	Iteração de um algoritmo de aprendizado
q	Número de neurônios de um mapa SOM
n_e	Número de épocas de treinamento
r_{bmu}	Posição do neurônio BMU no mapa
r_u	Posição do neurônio u no mapa
w_u	Vetor de peso de um neurônio
w_{bmu}	Vetores de pesos de neurônios vencedores
v	Vetor de dados de entrada
d_V	Camada de entrada de um SOM
d_A	Camada de saída de um SOM

1 Introdução

“Sigo sendo sempre a mesma
pessoa, mas creio ter mudado até
os ossos.” (*J. W. Goethe*)

A constante evolução tecnológica aliada à informatização dos mais variados processos organizacionais gerou um excesso de dados armazenados. Apesar disso, há pobreza de informação. Observa-se que a capacidade do ser humano em analisar conjuntos de dados de diferentes naturezas e em várias escalas é bem menor que sua capacidade de produzir conjuntos massivos de dados. Há, portanto, uma lacuna entre a produção de dados e o processamento analítico eficiente e de fácil interpretação, por analistas no domínio da aplicação.

Cada vez mais, busca-se identificar informações relevantes e transformá-las em conhecimento útil à tomada de decisões. Esta situação tem demandado a geração de novas técnicas, metodologias e ferramentas de análise de dados que assistam aos analistas humanos na compreensão de seus dados.

A análise exploratória de dados emprega diferentes técnicas de visualização e interpretação gráficas e quantitativas visando maximizar a obtenção de informações ocultas na estrutura dos conjuntos de dados. Dentre essas abordagens, estão os algoritmos de agrupamento de dados, um processo não-supervisionado de reconhecimento de padrões que cumpre um papel fundamental para a análise

exploratória de dados. Tais algoritmos trabalham na reunião de objetos que guardem significativa semelhança, o que facilita a descoberta de padrões, bem como a correta manipulação de outros objetos que venham juntar-se aos dados já coletados e agrupados.

As técnicas de agrupamento de dados, que não são comprobatórias e sim exploratórias por natureza, objetivam justamente realizar uma separação ótima entre os objetos de uma coleção. A questão central é como organizar os dados observados em estruturas que façam sentido, ou ainda, como desenvolver taxonomias capazes de separar os dados observados em diferentes grupos que ocorram naturalmente no próprio conjunto de dados.

1.1 Motivação

O agrupamento de dados, de aprendizagem não-supervisionada, é um problema difícil devido, por exemplo, à diversidade de formato dos dados, número de dimensões e grau de separabilidade. Para alcançar a solução de tal tarefa, existem na literatura centenas de algoritmos. No entanto, não existe um único algoritmo capaz de tratar adequadamente todos os tipos de dados. Uma dificuldade adicional está na validação dos resultados de agrupamento (particionamento dos dados), devido ao caráter inerentemente subjetivo do conceito de grupo.

Várias são as abordagens propostas à tarefa de agrupamento de dados, dentre elas, as Redes Neurais Artificiais (RNAs), que muito têm-se desenvolvido nos últimos tempos, com aplicações diversas. RNAs são sistemas paralelos, distribuídos e compostos por unidades de processamento simples - nodos - que calculam determinadas funções, normalmente não-lineares, com diferentes objetivos, como aproximar superfícies, representar modelos discriminatórios etc.

Dentre as RNAs, destaca-se neste trabalho a rede SOM (*Self-Organizing Maps*), a qual define “uma rede elástica” de pontos (parâmetros, referências ou vetores *codebook*) ajustados ao espaço de entrada dos dados para aproximar sua função de densidade de forma ordenada (KOHONEN, 2001). Sendo SOM uma abordagem paramétrica de análise de dados, a determinação de seus parâmetros pode influenciar no resultado final produzido pelo algoritmo, o que torna a sua modelagem um problema difícil.

As aplicações principais de SOM são visualização de dados complexos no formato bidimensional e criação de abstrações para agrupamento de dados. Contudo, para agrupamento de dados, SOM é comumente utilizada apenas como uma das etapas do processo, seguida da aplicação de técnicas clássicas de agrupamento de dados. Segundo (COTTRELL; BODT, 1996), a representação clássica de SOM não provê nenhuma informação sobre as distâncias locais entre neurônios adjacentes e, dois vizinhos podem estar muito longe entre si, no espaço da entrada. Portanto, não se tem idéia sobre eventuais grupos que possam existir em um mapa treinado. Dentro deste contexto, este trabalho explora uma metodologia que utiliza a RNA SOM de uma maneira alternativa, viável para aplicações de análise de agrupamento.

1.2 Objetivos

O objetivo principal do presente trabalho é explorar, de forma metodológica, a aplicação de SOM ao problema de análise de agrupamentos de dados numéricos, de maneira que se descubra a estrutura de grupos inerente aos conjuntos de dados analisados, caso haja uma. Este objetivo ramifica-se em:

1. A proposição de uma metodologia de agrupamento sobre o SOM que considere a topologia dos dados na descoberta de grupos;

2. A análise exploratória de dados em si, pela avaliação e interpretação dos agrupamentos obtidos.

1.3 Contribuições Esperadas

Ao alcançar o objetivo acima exposto, esta pesquisa contribui com a área de análise exploratória de dados nos seguintes quesitos:

- Apresentação de um novo processo de análise de agrupamento utilizando o algoritmo SOM;
- Desenvolvimento de um conjunto de considerações sobre a topologia dos dados e o uso do mapeamento do SOM no processo de agrupamento;
- Avaliação da metodologia proposta frente a trabalhos correlatos, de forma comparativa.

1.4 Organização do Trabalho

Além da introdução (Capítulo 1), que visa a contextualização do trabalho, os seguintes capítulos compõem este documento:

- O Capítulo 2 provê uma visão geral sobre análise de agrupamentos, situando o leitor nesta área de pesquisa.
- O Capítulo 3 versa sobre os Mapas Auto-Organizáveis de *Kohonen*, descrevendo suas principais características, questões sobre problemas de parametrização e métricas para aferir a qualidade de um mapa.
- A aplicação de SOM em análise exploratória (visualização e análise de agrupamento de dados) e trabalhos correlatos a esta pesquisa, que propõem

agrupamento de dados em duas fases, onde a primeira é o SOM, são abordados no Capítulo 4. Ainda neste capítulo, a descrição da metodologia de agrupamento proposta nesse trabalho é dada.

- Uma avaliação experimental realizada para averiguar a viabilidade e o potencial da proposta, bem como uma comparação com os resultados obtidos pela metodologia descrita em (VESANTO; ALHONIEMI, 2000) são apresentadas no Capítulo 5.
- As conclusões e trabalhos futuros dessa pesquisa são cobertos no Capítulo 6, que também enfatiza as principais contribuições nela alcançadas.
- No Anexo A constam os algoritmos de validação de agrupamentos citados nesta tese, seguido, por fim, da lista de referências bibliográficas utilizadas nesta pesquisa.

2 Análise de Agrupamento de Dados

“Fôssemos infinitos, tudo mudaria.
Como somos finitos, muito
permanece.” (*Bertold Brecht*)

O objetivo deste capítulo é descrever o processo de análise de agrupamentos, de forma a propiciar uma compreensão geral sobre o assunto. Além dos conceitos gerais do processo de análise de agrupamentos, encontram-se os seguintes tópicos: medidas utilizadas para estabelecer a proximidade (vizinhança) entre os dados, uma taxonomia para os principais métodos citados na literatura, uma introdução sobre validação de agrupamentos, e os desafios e tendências desta área.

2.1 Conceitos Gerais

Várias denominações são utilizadas para análise de agrupamentos, como *Data Clustering*, Taxonomia Numérica, Análise de Grupos ou Análise de Aglomerados (ou Conglomerados). De acordo com (EVERITT; LANDAU; MORVEN, 2001), análise de agrupamentos é provavelmente o nome mais genérico, preferido para referir-se a este processo, motivo pelo qual é adotado nesta tese.

Análise de Agrupamentos é o processo de agrupamento de dados, tal que objetos dentro de um grupo tenham alta similaridade, quando comparados a

outros objetos deste, e alta dissimilaridade a objetos de outros grupos. Para (CORMACK, 1971), a tarefa de agrupamento de dados está pautada em duas idéias básicas: a coesão interna dos objetos (homogeneidade) e o isolamento externo (separação) entre os grupos. Esta mesma definição é adotada por (HAIR et al., 1998) e (LATTIN; CARROLL; GREEN, 2003).

Para (EVERITT; LANDAU; MORVEN, 2001), a análise de agrupamentos é uma denominação genérica para uma ampla escala de métodos numéricos utilizados para examinar dados multivariados, visando encontrar conjuntos de observações homogêneas. Dada uma amostra de n dados (ou indivíduos), cada um deles medido segundo p variáveis, o objetivo é procurar um esquema que os agrupe em g grupos. Com este agrupamento, pode-se identificar relacionamentos úteis entre os dados, como similaridades e diferenças antes não reveladas.

É um processo de aprendizado não-supervisionado, uma vez que não há classes ou exemplos pré-definidos que evidenciem que algum tipo de relação deva ser válida entre os dados, ou mesmo a presença de tutores do domínio para supervisionar o aprendizado. Tanto o número ótimo de grupos quanto as características particulares que revelam semelhanças (ou diferenças) devem ser determinados pelo próprio processo.

Agrupamento de dados, que engloba uma variedade de conceitos e algoritmos, tem suas origens em várias áreas, incluindo Inteligência Artificial, Estatística e Banco de Dados. Atualmente é considerada uma das tarefas mais usuais do processo de mineração de dados (*Data Mining* - DM) para descobrir grupos e identificar distribuições e padrões interessantes que estão subjacentes aos dados (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001).

Um sistema de aprendizado não-supervisionado usa análises de redundância da informação recebida para a aquisição do conhecimento, porque não existe

- ou não se assume, um conjunto de dados para testes de classificação. Essa redundância pode ser obtida pelo cálculo de propriedades estatísticas do conjunto de dados como média, variância e covariância, ou aferindo a forma como os dados podem estar agrupados.

Na maioria das aplicações de análise de agrupamentos busca-se um bom particionamento dos dados, de forma que cada objeto pertença a um único grupo, e o conjunto completo dos grupos contenha todos os objetos. Em algumas circunstâncias, porém, conjuntos sobrepostos também podem constituir uma solução aceitável.

Uma proposta de divisão da análise de agrupamentos em etapas é dada por (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001), na Figura 2.1, que ilustra os passos básicos dessa tarefa, brevemente explicados a seguir.

Figura 2.1: Etapas de um processo de análise de agrupamentos.
(HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001)

- **Seleção de Atributos:** Nesta fase define-se quais atributos serão utilizados para o agrupamento; por vezes, um pré-processamento se faz necessário.
- **Escolha do algoritmo de Agrupamento:** Envolve a escolha de um algoritmo que defina um “bom” esquema de agrupamento para um conjunto de dados. Escolher um algoritmo de agrupamento envolve definir uma medida de proximidade e uma função critério, fatores esses decisivos na definição de uma estrutura de grupos que realmente represente o conjunto de dados.
- **Validação dos Resultados:** Etapa em que a qualidade da estrutura de grupos gerada é avaliada. Uma vez que os algoritmos definem os grupos, uma avaliação se faz necessária para verificar a correção dos resultados.
- **Interpretação dos Resultados:** Neste ponto há a interpretação da estrutura de grupos gerada, por parte de um especialista de domínio. Pode-se também integrar os resultados de agrupamento com outras análises e evidências experimentais para melhores conclusões.

Não há, contudo, consenso entre os autores da área sobre a divisão em etapas da análise de agrupamento de dados. Alguns autores (JAIN; DUBES, 1988), (THEODORIDIS; KOUTROUMBAS, 2005), (PERES, 2006) incluem neste processo uma fase de análise de tendência a agrupamentos (*Cluster Tendency*). Esta etapa consiste em testes que indicam se os dados disponíveis possuem ou não uma estrutura de agrupamento. É uma forma de prevenir a má interpretação dos dados, pois dependendo do algoritmo utilizado, uma estrutura será imposta, mesmo que os dados sejam de natureza completamente aleatória, o que não fará sentido.

A análise de agrupamentos é um problema difícil porque muitos fatores críticos não estão incluídos no problema dado, como boas medidas de proximidade, definição de funções critério, algoritmos apropriados e suas condições iniciais.

Além disso, sabe-se que nenhum método de agrupamento pode tratar adequadamente todos os tipos de estruturas de grupos (com diferentes formas, tamanhos e densidades) (KOTSIANTIS; PINTELAS, 2004). A Figura 2.2 ilustra diferentes formatos de grupos, que podem ser favorecidos ou prejudicados na identificação, de acordo com o método adotado.

Figura 2.2: Exemplos de geometrias de grupos. (a) Grupos compactos. (b) Grupos alongados. (c) Grupo esférico e elipsoidal.
(THEODORIDIS; KOUTROUMBAS, 2005)

Ademais, em muitas aplicações a noção de um grupo não é bem definida. A Figura 2.3 traz um conjunto de vinte pontos e três diferentes partições. A definição de qual é a divisão correta só pode ser feita a partir de uma análise mais cuidadosa sobre os dados, com auxílio de um especialista no domínio da aplicação. Essa figura evidencia que a definição de um conjunto é imprecisa e que a melhor definição de grupo depende da natureza dos dados e dos resultados desejados, ou seja, a subjetividade é uma constante.

Uma fase bastante importante da análise de agrupamentos é a fixação dos critérios de homogeneidade ou função critério. Critérios distintos levam a grupos homogêneos distintos e o tipo de homogeneidade depende dos objetivos a serem alcançados (BUSSAB; MIAZAKI; ANDRADE, 1990).

Vários são os problemas enfrentados pelos algoritmos de aprendizado não-supervisionado. Para (HAN; KAMBER, 2001), são necessidades típicas em um

Figura 2.3: Diferentes grupos formados sobre um mesmo conjunto de dados.
(KUMAR; TAN; STEINBACH, 2005)

processo de análise de agrupamento de dados:

- Escalabilidade;
- Habilidade em manipular diferentes tipos de dados;
- Descoberta de grupos com formas arbitrárias;
- Exigências mínimas do conhecimento do domínio para determinar os parâmetros de entrada;
- Capacidade de tratar dados com ruído;
- Insensibilidade à ordem dos registros de entrada;
- Alta dimensionalidade;
- Permitir a incorporação de restrições específicas do usuário;
- Interpretabilidade e usabilidade.

Cabe salientar que os algoritmos existentes não dão conta, simultaneamente, de todos esses requisitos de forma efetiva.

2.2 Medidas de Proximidade

Na formação dos agrupamentos calcula-se a proximidade entre os dados, tanto da magnitude de distância entre dois objetos de um mesmo grupo (intra-grupos) quanto da magnitude das distâncias entre grupos (inter-grupos).

Medidas de proximidade intra-grupo determinam a similaridade entre exemplares pertencentes a um mesmo grupo - medem a densidade das conexões entre instâncias de um mesmo grupo, ao passo que as medidas inter-grupos determinam a similaridade entre agrupamentos, ou seja, a conectividade entre grupos distintos. A medida intra-grupo alta indica um bom arranjo de agrupamentos, pois as instâncias em um mesmo grupo são altamente relacionadas umas às outras. Uma medida inter-grupo baixa é desejável porque indica que os grupos individuais são, na maior parte, independentes entre si.

2.2.1 Proximidade entre dois Objetos

Como os objetos submetidos ao processo de particionamento em grupos são caracterizados por vetores de dados, a similaridade entre dois objetos corresponde a uma métrica que compara a igualdade desses vetores de dados. Ainda, pode-se calcular a dissimilaridade entre esses vetores medindo a distinção entre os objetos, ou seja, a distância entre eles.

Formalmente, de acordo com (THEODORIDIS; KOUTROUMBAS, 2005), uma medida de dissimilaridade d em um conjunto de dados X é uma função:

$$d : X \times X \rightarrow \mathbb{R} \quad (2.1)$$

onde $\mathfrak{R}$ é o conjunto dos números reais, tal que

$$\exists d_0 \in \mathfrak{R} : -\infty < d_0 \leq d(x, y) < +\infty, \forall x, y \in X \quad (2.2)$$

$$d(x, x) = d_0, \forall x \in X \quad (2.3)$$

e

$$d(x, y) = d(y, x), \forall x, y \in X \quad (2.4)$$

Ainda,

$$d(x, y) = d_0, \text{ se e somente se } x = y \quad (2.5)$$

e

$$d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in X \quad (2.6)$$

A Equação 2.6 é conhecida também como desigualdade triangular.

Uma métrica de similaridade s em X (um conceito simetricamente oposto à dissimilaridade), é definida como

$$s : X \times X \rightarrow \mathfrak{R} \quad (2.7)$$

tal que

$$\exists s_0 \in \mathfrak{R} : -\infty < s(x, y) \leq s_0 < +\infty, \forall x, y \in X \quad (2.8)$$

Ainda,

$$s(x, y) = s_0, \text{ se e somente se } x = y \quad (2.9)$$

e

$$s(x, y)s(y, z) \leq [s(x, y) + s(y, z)]s(x, z), \forall x, y, z \in X \quad (2.10)$$

As medidas de proximidade¹ mais utilizadas em agrupamento de dados são:

1. Distância Euclidiana, que é dada por:

$$d_e(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} \quad (2.11)$$

2. Métrica de *Minkowsky*, ou norma L_λ , que é uma generalização da Equação 2.11, para $\lambda > 2$:

$$d_\lambda(x, y) = \sqrt[\lambda]{\sum_{i=1}^p |x_i - y_i|^\lambda} \quad (2.12)$$

3. Distância *City-block*, um caso especial da Equação 2.12, quando $\lambda = 1$:

$$d_{cb}(x, y) = \sum_{i=1}^p |x_i - y_i| \quad (2.13)$$

4. Métrica de *Tanimoto*:

$$d_T(x, y) = \frac{x^T y}{x^T x + y^T y - x^T y} \quad (2.14)$$

5. Métrica de *Mahalanobis*:

$$d_M(x, y) = \sqrt{(x - y)^T cov^{-1}(x - y)} \quad (2.15)$$

onde *cov* é uma matriz de covariância simétrica positiva.

Embora a distância euclidiana seja a medida de proximidade mais usada pelos algoritmos de agrupamento, há várias outras medidas utilizadas para este fim. Qual dessas medidas utilizar é uma decisão de projeto, intrinsecamente ligada à caracterização dos dados.

¹Para uma discussão aprofundada sobre medidas de proximidade, (EVERITT; RABEHESKETH, 1997) é uma bibliografia indicada.

2.2.2 Proximidade entre dois Grupos

A maioria das funções de proximidade entre grupos (d_g) é baseada em medidas de proximidade, d , entre vetores (THEODORIDIS; KOUTROUMBAS, 2005). Sejam D_i , D_j dois grupos de vetores de dados, as funções mais comuns são:

- **Proximidade Mínima (*Single Link*):** É definida como a mínima distância (máximo de similaridade) entre dois pontos em dois grupos diferentes.

$$d_{g_min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} d(x, y) \quad (2.16)$$

- **Proximidade Média (*Average Link*):** Calcula-se a proximidade média entre pares de cada grupo, para todos os pares de pontos de grupos diferentes.

$$d_{g_avg}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{x \in D_i} \sum_{y \in D_j} d(x, y) \quad (2.17)$$

onde n_{D_i} e n_{D_j} são as cardinalidades de D_i e D_j , respectivamente.

- **Proximidade Máxima (*Complete Link*):** A proximidade de dois grupos é definida como a máxima distância (mínimo de similaridade) entre quaisquer dois pontos em dois grupos diferentes.

$$d_{g_max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} d(x, y) \quad (2.18)$$

- **Proximidade pela Mínima Variância (Método *Ward*):** A proximidade entre dois grupos é definida em termos da soma dos quadrados dos erros de variância mínima (EL-HAMDOUCHI; WILLETT, 1986).

$$d_{g_ward}(D_i, D_j) = \frac{n_{D_i} n_{D_j} d_e(D_i, D_j)}{n_{D_i} + n_{D_j}} \quad (2.19)$$

onde n_{D_i} e n_{D_j} são as cardinalidades de D_i e D_j , respectivamente, e $d_e(D_i, D_j)$ a distância Euclidiana.

A Figura 2.4 exemplifica as três primeiras definições de proximidade entre grupos supracitadas.

Figura 2.4: Definição visual de proximidade entre grupos.

A aplicação de um método ou outro varia de acordo com o formato dos grupos. Por exemplo, (BAEZA-YATES; FRANKS, 1992) afirmam que *Complete link* é melhor para grupos mais compactos, enquanto que em (ROCHA; LACHI, 2005) é dito que *Single link* produz grupos mais alongados.

2.2.3 Proximidade intra Grupos

O objetivo dessas funções é medir a proximidade entre os pontos de um agrupamento. Seja D_k , um grupo de dados, são exemplos dessas funções:

- **Vizinho mais próximo:** É definida como a mínima distância entre dois pontos em um mesmo grupo.

$$d_{i_nn}(D_k) = \min_{x,y \in D_k} d(x,y) \quad (2.20)$$

- **Distância do Centróide:** É definida como a distância média entre os pontos e o centro de um grupo de dados.

$$d_{i_cen}(D_k) = \frac{1}{N_k} \sum_{x \in D_k} d(x, c_k) \quad (2.21)$$

onde N_k representa o número de elementos do grupo D_k e c_k corresponde ao centro do grupo.

2.3 Categorias de Algoritmos de Agrupamento

Existe na literatura vários algoritmos para agrupamento de dados, que utilizam maneiras diferentes para a identificação e representação de agrupamentos. A escolha por aplicar um determinado algoritmo depende do tipo de dados que se pretende explorar, da aplicação e de seus objetivos particulares. Ao usar análise de agrupamentos como uma ferramenta descritiva ou exploratória, é possível optar pela aplicação de vários algoritmos sobre o mesmo conjunto de dados, e averiguar o que esses dados podem revelar.

As técnicas de agrupamento podem ser organizadas nas seguintes categorias:

- **Métodos baseados em Particionamento:** Dada uma base de dados com n objetos e k grupos a formar, um algoritmo de particionamento organiza os objetos em k partes ($k \leq n$), onde cada parte representa um grupo.

Os grupos são construídos de forma a otimizar um critério de particionamento objetivo, comumente chamado de função de similaridade, tal como uma medida de distância. Assim, os objetos alocados dentro de um grupo serão “similares” e os objetos alocados em grupos diferentes serão “dissimilares”, com relação aos atributos que formam a base de dados.

Esses métodos são divididos em duas grandes categorias, os algoritmos baseados em centróides e os baseados em medóides. Os algoritmos baseados em centróides, como o *k-means*, usam o centro geométrico de cada grupo para representá-lo. O algoritmo *k-medoids* utiliza o medóide - que é o objeto localizado mais próximo ao centro do grupo - para representar cada grupo, tentando amenizar a influência de pontos isolados a partir da minimização da dissimilaridade entre os dados de um grupo e o seu ponto de referência.

O Algoritmo *k-means* (Algoritmo 1), é o mais conhecido deste grupo (JAIN;

MURTY; FLYNN, 1999), e, de acordo com (BERKHIN, 2002), é a ferramenta de agrupamento mais utilizada tanto em aplicações científicas quanto na indústria.

Algorithm 1 - Algoritmo *k-means*

Entrada: Conjunto de dados e um valor para k .

Saída: Conjunto de dados agrupado.

Selecione k pontos, aleatoriamente, como centróides iniciais.

repeat

 Atribua cada ponto ao centróide mais próximo;

 Recalcule o centróide para cada grupo;

until Estabilizar

Este algoritmo funciona bem apenas para grupos esféricos, com resultado fortemente dependente dos centróides iniciais. É sensível a dados espúrios (*outliers*) e exige conhecimento prévio do número de grupos existentes.

- **Métodos Hierárquicos:** Um método de agrupamento hierárquico trabalha com agrupamento de dados em uma árvore de grupos (dendrograma). Podem ser classificados em agrupamento hierárquico aglomerativo ou divisivo, dependendo de como a decomposição hierárquica é formada, se, respectivamente, de forma *bottom-up* ou *top-down* (JAIN; DUBES, 1988).

O algoritmo aglomerativo inicia com cada elemento pertencendo a um grupo. A partir daí encontra dois grupos mais similares e forma um novo grupo. Essa operação é então repetida, considerando os novos grupos, até haver apenas um grupo. Já na abordagem divisiva, o processo inicia-se com apenas um agrupamento contendo todos os dados e segue dividindo-o recursivamente, segundo alguma métrica, até que alcance um critério de parada, freqüentemente, o número de grupos desejados.

Algoritmos hierárquicos são utilizados quando há necessidade de gerar taxonomias ou hierarquias entre os dados. A Figura 2.5 ilustra um agrupamento hierárquico aglomerativo *Single link* (Equação 2.16) no espaço

euclidiano com seis pontos, com representações distintas.

Figura 2.5: Exemplo de um agrupamento hierárquico aglomerativo.

Os métodos hierárquicos, embora simples, sofrem de inabilidade em executar ajustes uma vez que uma decisão de junção ou divisão tenham sido executadas. Esse é um fato crítico, pois uma vez que um grupo de dados é unido ou particionado, a próxima iteração será processada com base nos grupos recém-formados, o que pode comprometer a qualidade do resultado final. Além disso, podem ser problemáticos ao lidar com conjuntos de dados de alta dimensionalidade ou com ruídos.

- **Métodos baseados em Densidade:** Baseiam-se no conceito de densidade, onde os grupos crescem de acordo com a densidade da vizinhança dos objetos ou de acordo com alguma função de densidade especificada. Assim, para cada objeto em um grupo, a vizinhança deste, dentro de um raio dado, deve conter ao menos um número mínimo de outros objetos. Os agrupamentos são definidos matematicamente pela identificação dos pontos mais densos (HAN; KAMBER, 2001).

Para descobrir grupos de formato arbitrário, métodos de agrupamento baseados em densidade têm sido desenvolvidos, que tipicamente consideram gru-

pos como sendo as regiões densas de objetos em dados esparsos, separados por regiões de baixa densidade, que representam ruídos.

- **Métodos baseados em Grades:** Essa abordagem usa uma estrutura de dados em grade de multiresolução, a qual quantiza o espaço em um número finito de células, que formam a estrutura em grade onde todas as operações de agrupamento são realizadas. O princípio básico desses métodos é primeiro sumariar o conjunto de dados com uma representação de grade, e então unir as células da grade para obter os grupos. Esta abordagem particiona o espaço dos dados em células disjuntas na grade. Há uma probabilidade elevada de que todos os pontos pertencentes a um mesmo grupo estejam em uma mesma célula da grade. Dessa forma, todos os pontos pertencentes a uma mesma célula podem ser agregados e tratados como um único objeto (QIU; ZHANG; SHEN, 2005). A principal vantagem dessa abordagem é a rapidez no tempo de processamento, que geralmente é independente do número de objetos de dados; contudo, é dependente do número de células em cada dimensão no espaço quantizado.
- **Métodos baseados em Modelos:** Os métodos de agrupamento baseados em modelos tentam otimizar o ajuste entre os dados de entrada e algum modelo matemático. Nestes métodos, um modelo hipotético é criado para cada grupo, e encontra-se o melhor ajuste do objeto ao modelo dado. Pode-se localizar grupos por meio de uma função densidade que reflita a distribuição espacial dos objetos. Tais métodos geralmente são baseados na suposição de que os dados são gerados por uma mistura de distribuições de probabilidade subjacentes. Duas abordagens desses métodos são:

Agrupamento conceitual: É uma forma de agrupamento em aprendizagem de máquina que, dado um conjunto de objetos não nominados, produz um esquema de classificação sobre eles (BISWAS; WEINBERG; FISHER,

1998). Diferentemente das abordagens de agrupamento convencionais, além de identificar grupos de objetos semelhantes, o agrupamento conceitual gera também conceitos/descrições de cada grupo encontrado. É, portanto, um processo em duas fases: primeiro é realizado o particionamento e depois, a caracterização dos dados.

Redes Neurais Artificiais: Esta abordagem tenta representar cada grupo por meio de um exemplar, que atua como um “protótipo” de grupo e que não necessariamente corresponde a um dos objetos ou vetor de dados particular. Novos objetos podem ser atribuídos ao grupo cujo protótipo lhe seja mais próximo, baseado em alguma medida de distância.

A Tabela 2.1, elaborada principalmente a partir de leituras em (JAIN; MURTY; FLYNN, 1999), (BERKHIN, 2002), (KOTSIANTIS; PINTELAS, 2004), (GRIRA; CRUCIANU; BOUJEMAA, 2004) e (XU; WUNSCH, 2005), apresenta, a título de ilustração, alguns exemplos de algoritmos facilmente encontrados na literatura, classificados na taxonomia acima exposta. SOM somente pode ser incluído nesta tabela se o número de neurônios do mapa for considerado igual ao número de grupos desejado.

CLASSE		Exemplos de Algoritmos
Particionamento		<i>k-means</i>
		<i>k-medoids</i>
		CLARA
		CLARANS
		CLUSTER
		PAM
Hierárquicos		AGNES
		DIANA
		CURE
		CHAMELEON
		BIRCH
		ROCK
		CLINK
Densidade		DBSCAN
		OPTICS
		DENCLUE
		GDBSCAN
		DBCLASD
Grades		STING
		CLIQUE
		WAVECLUSTER
		BANG-CLUSTERING
		GRIDCLUST
Modelos	Estatísticos	COWEB
		CLASSIT
		AUTOCLASS
		SNOB
		MCLUST
	Redes Neurais	SOM
		ART

Tabela 2.1: Exemplos de algoritmos de agrupamento, por categorias.

Embora o foco deste trabalho seja agrupamento de dados numéricos, nem sempre os conjuntos de dados disponíveis possuem apenas atributos numéricos. Ao contrário, muitos incluem também dados categóricos relevantes à classificação, comuns em várias áreas, como nas ciências sociais e biomédicas.

Dados categóricos são aqueles cujas variáveis consistem em uma série de estados discretos, tal que cada objeto pertence a apenas um dos estados, chamados de:

- **nominais**, quando não existe uma ordem específica entre os estados diferentes, por exemplo, a cor azul ou preta para o atributo cor dos olhos, a afiliação religiosa (católico, judeu, protestante, espírita, budista, outra), o sexo, ou o estado civil de um indivíduo. A ordenação, inclusive, é irrelevante para análises estatísticas;
- **ordinais**, quando existe uma ordem subjacente que pode ser imposta aos dados, como no atributo hierarquia de cargos “diretor > coordenador > operador”, ou o resultado de avaliação de um artigo submetido a uma conferência, que pode ser “Fortemente Aceito”, “Fracamente aceito”, “Fracamente Rejeitado” e “Fortemente Rejeitado”. Apesar dessa ordenação em categorias, a distância absoluta entre elas é desconhecida. Neste último exemplo, pode-se concluir que um artigo avaliado como “Fracamente Aceito” é melhor que um artigo tido como “Fracamente Rejeitado”. Contudo, não há um valor numérico que indique esse grau de qualidade.

Alguns dados categóricos podem ser - e em geral são - representados por valores numéricos. Em (THEODORIDIS; KOUTROUMBAS, 2005), há dois exemplos desse fato: o atributo sexo, onde é atribuído valor 1 para masculino e 0 para feminino, e o atributo desempenho acadêmico, com valores 4, 3, 2 e 1 correspondendo, respectivamente, a “Excelente”, “Ótimo”, “Bom” e “Insuficiente”. Entretanto, com-

parações quantitativas nesses atributos não serão significativas.

No caso dos dados nominais, não se pode estabelecer uma função de distância natural entre dois elementos. Um exemplo de uma base de dados com informações sobre carros, com os atributos fabricante, modelo, preço, cor e cliente é apresentado por (GIBSON; KLEINBERG; RAGHAVAN, 2000). O fabricante é um típico exemplo de dado categórico nominal, sendo difícil estabelecer uma relação de proximidade entre os dados. Por exemplo, “Honda” e “Toyota” podem ser considerados próximos se for considerada a origem da fábrica, mas podem estar distantes na faixa de preço médio de seus modelos. Desta forma, não se pode usar métricas tradicionais para encontrar a proximidade, como aquelas da Seção 2.2, sendo necessária outra abordagem para lidar com estes dados.

Vários trabalhos têm sido desenvolvidos nesse sentido. Métodos estatísticos para determinar a associação de variáveis categóricas (seus relacionamentos) são apresentados por (SCHEAFFER, 1999) e (AGRESTI, 2002). O trabalho de (CHEN; LIU, 2003) faz um apanhado sobre técnicas de agrupamento para dados categóricos. Uma forma de medir uma distância numérica entre atributos categóricos, de forma a utilizar algoritmos como *k-means* ou hierárquicos, é apresentada por (PALMER; FALOUTSOS, 2003). Outras abordagens, como (GIBSON; KLEINBERG; RAGHAVAN, 2000) e (ZHANG et al., 2000), buscam criar algoritmos que trabalhem a partir dos relacionamentos dos dados categóricos de uma base de dados.

Há ainda outras abordagens para agrupamento de dados, como as baseadas em Algoritmos Genéticos (JAIN; MURTY; FLYNN, 1999) e as baseadas em Máquinas de Vetores Suporte (*Support Vector Machines* - SVMs) (VAPNIK et al., 2001).

Tanto os algoritmos acima descritos quanto a abordagem desta tese são *crisp*, ou seja, a noção de pertinência é categórica: elementos pertencem ou não pertencem a um dado agrupamento. Para algoritmos de agrupamento *fuzzy*, que

permitem categorias sobrepostas para os dados (grupos mal delineados), indica-se (HÖPPNER et al., 1999).

2.4 Validação de Agrupamentos

Em classificação supervisionada de dados, a avaliação do classificador é naturalmente vista como parte integrante do processo, com estratégias bem definidas. O mesmo, no entanto, não ocorre no aprendizado não-supervisionado, onde a avaliação - ou validação como é comumente chamada, não é tão bem desenvolvida ou aplicada.

Dado um conjunto de dados, um algoritmo de agrupamento sempre irá gerar uma divisão, sem se preocupar se esta estrutura realmente existe ou não. É importante saber se esses grupos estavam realmente ocultos aos dados ou se são apenas artefatos, sem significado, gerados pelo algoritmo.

As várias abordagens existentes para análise de agrupamentos podem gerar grupos diferentes para os mesmos dados. Ainda, para um mesmo algoritmo, a mudança de parâmetros ou mesmo a ordem de apresentação dos padrões de entrada podem afetar o resultado final. Portanto, uma boa avaliação é fundamental para aumentar o grau de confiabilidade nos resultados gerados pelo algoritmo por parte dos especialistas de domínio.

Distinguir se há uma estrutura não aleatória nos dados é apenas um item importante da validação de agrupamentos. Outros aspectos também relevantes são citados por (KUMAR; TAN; STEINBACH, 2005):

1. Determinar o número “correto” de grupos;
2. Avaliar o quanto o resultado de uma análise de agrupamento se ajusta aos dados, sem referência de informações externas;

3. Comparar os resultados da análise de agrupamento com resultados externos conhecidos, como rótulos de classes fornecidos;
4. Comparar dois conjuntos de grupos para determinar qual é o melhor.

Para responder às questões supracitadas, vários índices - ou medidas de avaliação - são aplicados para testar os agrupamentos. Há três categorias de testes utilizados para validação (JAIN; MURTY; FLYNN, 1999), (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a), (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002b), (KUMAR; TAN; STEINBACH, 2005):

- **Índices Internos (não-supervisionados):** Não dependem de conhecimento prévio. Examinam diretamente a estrutura de agrupamento a partir dos dados originais. Geralmente estes índices são divididos em medidas de coesão, que determinam o quão relacionados os objetos de um conjunto estão, e em medidas de separação, que dizem o quão bem separado um grupo está de outro. Índice Dunn, Silhueta e Davies-Bouldin são exemplos dessa abordagem.
- **Índices Externos (supervisionados):** São baseados em alguma estrutura pré-especificada, a partir de algum conhecimento *a priori* sobre os dados, como um padrão de validação à solução apresentada. Se existem informações externas sobre os dados, por exemplo, rótulos de classes para os objetos de dados, o procedimento usual é medir o grau de correspondência entre os rótulos dos grupos gerados e os rótulos das classes.

De acordo com (KUMAR; TAN; STEINBACH, 2005), essas medidas são classificadas em:

- Medidas orientadas a classificação, que avaliam se os objetos de um

grupo são todos do mesmo grupo. Entropia, pureza, precisão, cobertura (*recall*), medida F , são exemplos destas técnicas.

- Medidas orientadas a similaridade, que medem se dois objetos pertencentes a uma mesma classe estão em um mesmo grupo e vice-versa. Coeficiente de *Jaccard* é um exemplo desta abordagem.

- **Índices Relativos:** Buscam, por meio da comparação de diferentes estruturas de agrupamentos, dar uma referência de qual delas melhor revela as características reais de seus objetos. Baseiam-se em avaliar a consistência ou estabilidade dos algoritmos de agrupamento, comparando os grupos obtidos pelo mesmo algoritmo sob diferentes condições.

O desempenho dos índices de validação é bastante variável, mas fornecem mais um instrumento de apoio à confiabilidade dos grupos gerados.

Há alguns trabalhos na literatura, a exemplo de (MILLIGAN; COOPER, 1985) e de (BRUN et al., 2007), que analisam, de forma experimental e comparativa, várias medidas de avaliação de agrupamentos, no intuito de revelar a efetividade dessas medidas, o que se configura como resultados pontuais, fortemente ligados às condições das simulações realizadas, como a distribuição dos conjuntos de dados, os algoritmos de agrupamento e a parametrização adotados.

2.5 Considerações Finais

A análise de agrupamentos é uma importante atividade humana que, por ajudar na organização de tarefas e na compreensão do mundo, está presente no cotidiano desde os primeiros anos.

Com o advento da mineração de dados, essa tarefa ganhou novo impulso na pesquisa por novas abordagens, e tem sido usada em inúmeras aplicações, como

reconhecimento de padrões, análise de dados e processamento de imagens.

Segundo (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a), se valores impróprios forem atribuídos aos parâmetros dos algoritmos, isso resultará em um esquema de particionamento que não será ótimo para o conjunto de dados, conduzindo a decisões erradas. Vários são os esforços em auxiliar a tomada de decisão sobre qual é a melhor estrutura de agrupamento que representa um conjunto de dados, uma vez que não há um critério padrão para guiar o projetista na seleção de atributos ou de algoritmos. Os critérios de validação provêm algumas idéias sobre a qualidade da solução de agrupamento, porém, a seleção de critérios ótimos de validação ainda é um vasto campo para pesquisas.

Neste capítulo, uma abordagem geral sobre a análise de agrupamentos foi dada, sem a pretensão de ser exaustiva. Os conteúdos aqui tratados serão úteis ao restante da leitura deste trabalho.

3 Mapas Auto-Organizáveis

“Aprender é a única coisa de que a mente nunca se cansa, nunca tem medo e nunca se arrepende.”

(Leonardo da Vinci)

O objetivo deste capítulo é discutir princípios de auto-organização, introduzir os conceitos principais de uma RNA SOM e apresentar medidas de qualidade comumente utilizadas para avaliação de um mapa auto-organizado.

3.1 Redes Neurais Artificiais e Princípios de Auto-organização

As redes neurais artificiais (RNAs) foram inspiradas nas redes neurais biológicas, de forma que o conhecimento seja adquirido pela rede por um processo de aprendizagem, e que as forças de conexão entre neurônios (os pesos sinápticos) sejam utilizadas para armazenamento dos conceitos aprendidos. Esta seção traz uma breve explanação sobre RNA. Para maiores informações sobre o assunto, sugere-se referências como (HAYKIN, 1999) e (FAUSSET, 1994).

De acordo com (HAYKIN, 1999) e (BRAGA; CARVALHO; LUDERMIR, 2000), uma rede neural é um processador maciçamente paralelamente distribuído, constituído de unidades de processamento simples, que têm a propensão natural

para armazenar conhecimento experimental e torná-lo disponível para o uso. Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais. Estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede.

A definição da arquitetura de uma RNA é de suma importância na sua concepção, independente de sua categoria, seja ela Rede Perceptron com uma Única Camada, Rede Perceptron Multicamadas, Redes de Função de Base Radial, Redes Competitivas, Mapas Auto-Organizáveis, Rede de *Hopfield* ou Redes ART (*Adaptive Resonance Theory*), pois está intrinsecamente ligada ao problema ao qual a rede será submetida.

Uma rede neural é caracterizada pela sua topologia, pelo seu número de camadas, pela força de conexão entre pares de neurônios (pesos sinápticos), pelas propriedades dos neurônios, e pela forma de atualização dos vetores de pesos (PAL; SRIMANI, 1996).

Dentre as várias propriedades e capacidades de uma RNA está a auto-organização. Segundo (HAYKIN, 1999), o princípio de auto-organização é a possibilidade de muitas interações locais entre neurônios vizinhos, originalmente aleatórias, poderem se fundir em estados de ordem global e finalmente levar a padrões espaciais ou ritmos temporais. Segundo (MALSBERG, 1990), este fato é de extrema importância em interações locais para criar estados de ordem global, conduzindo a um comportamento coerente.

O postulado de Hebb (HEBB, 1949), denominado plasticidade Hebbiana, diz que “quando um axônio¹ de uma célula A está próximo o suficiente para exci-

¹Axônio é o prolongamento dos neurônios responsável pela condução dos impulsos elétricos que partem do corpo celular até outros locais mais distantes do sistema nervoso (BEAR; CONNORS; PARADISO, 2002).

tar uma célula B e repetidamente ou persistentemente participa de sua ativação, algum processo de crescimento ou alteração metabólica acontece em uma ou ambas as células, tal que a eficiência de A como uma das células que ativa B é aumentada”.

Expandindo essa afirmação, pode-se dizer que, se a ativação de um neurônio tiver grande influência na ativação de um outro neurônio, a ligação entre estes deve ser reforçada; ao contrário, se a ativação de um neurônio não influencia a ativação de um segundo neurônio, a ligação entre estes dois neurônios deve ser enfraquecida ou até mesmo removida. Portanto, pode-se assumir que o conceito de plasticidade está relacionado à auto-organização.

O desenvolvimento de mapas auto-organizáveis como modelo neural artificial foi motivado por uma característica do cérebro humano, que é a sua organização em várias regiões, de forma tal que entradas sensoriais distintas sejam representadas por mapas computacionais topologicamente ordenados. Como exemplo, cita-se as entradas sensoriais táteis, visuais, acústicas, que são mapeadas em diferentes áreas do córtex cerebral, de uma forma topologicamente ordenada (BEAR; CONNORS; PARADISO, 2002).

Assim, pode-se considerar que um mapa constitui um bloco básico de construção na estrutura de processamento de informação do sistema nervoso. Um mapa computacional é definido como um arranjo de neurônios que operam nos sinais que transportam informação sensorial em paralelo. Conseqüentemente, os neurônios transformam sinais de entrada em uma distribuição de probabilidade, a qual é codificada na região que representa a máxima atividade relativa dentro do mapa.

Em (MALSBERG, 1990), três princípios de auto-organização aplicados ao cérebro e às redes neurais artificiais são apresentados:

- **Princípio 1:** Modificações dos pesos sinápticos tendem a se auto-amplificar. Este princípio está intrinsecamente ligado à formulação de Hebb supra, e indica que a organização ocorre em dois níveis, de atividade e de conectividade.
- **Princípio 2:** A limitação de recursos leva à competição entre sinapses, e com isso, à seleção das sinapses que crescem mais vigorosamente às custas de outras. Como os recursos do sistema devem ser compartilhados, somente as sinapses que se auto-fortalecem tendem a permanecer no sistema, e cada vez mais fortes. Este princípio é possibilitado pela plasticidade sináptica.
- **Princípio 3:** As modificações em pesos sinápticos tendem a cooperar. A presença de uma sinapse forte pode reforçar o ajuste de outras sinapses, apesar do ambiente global competitivo da rede.

Um quarto princípio de auto-organização é definido por (BARLOW, 1989):

- **Princípio 4:** Ordem e estrutura nos padrões de informação representam informação redundante, adquirida pela rede neural na forma de conhecimento, que é um pré-requisito necessário à aprendizagem auto-organizada. Este princípio diz respeito às condições dos estímulos do sistema neural, bem como o fato de que a redundância de informação é utilizada na definição de padrões úteis.

3.2 Conceitos Básicos do SOM

Um Mapa Auto-organizável (*Self-Organizing Map* ou SOM, como é conhecido) é uma arquitetura de rede neural, com aprendizado não-supervisionado e competitivo, proposta por *Teuvo Kohonen* (KOHONEN, 1989). Possui a habilidade de

formar mapeamentos que preservam a topologia entre os espaços de entrada (dos dados) e de saída (da rede, ou de representação).

No aprendizado competitivo os neurônios competem entre si para responder a um estímulo. Esse processo, segundo (RUMELHART; ZIPSER, 1985), é constituído por:

- Um conjunto de neurônios diferenciados por seus pesos, que respondem de forma diferente a um determinado conjunto de estímulos;
- Um limiar imposto sobre a força de cada neurônio;
- Um mecanismo de competição entre os neurônios, onde apenas um, entre um grupo de neurônios, possa responder a uma classe de estímulos de entrada.

Dadas as características acima, a rede SOM é baseada em um mapa de neurônios, cujos pesos são adaptados às entradas de vetores semelhantes presentes em um conjunto de treinamento (KOHONEN, 2001). Esse mapa pode ser estruturado de forma hexagonal ou retangular, como visto na Figura 3.1. Os polígonos na figura representam três diferentes níveis de vizinhança no mapa. A forma de ligação entre os neurônios do mapa define a vizinhança do neurônio, que será de 4 neurônios se o mapa for retangular ou com 6 vizinhos se o formato for hexagonal.

Figura 3.1: Tipos de estrutura de um mapa auto-organizável.

O aprendizado competitivo, sob o prisma de redes neurais artificiais, diz respeito à quantização vetorial, a ser descrita na Subseção 3.4.1, e três processos são essenciais à formação do mapa (HAYKIN, 1999):

- **Competição:** Para cada padrão de entrada, os neurônios do mapa competem entre si para representá-lo. Para cada neurônio é computado seu respectivo valor, a partir de uma função de proximidade ao dado de entrada. O neurônio cujo vetor de pesos for mais próximo do dado, segundo a métrica calculada, é declarado BMU (*Best Matching Unit*), ou neurônio vencedor.
- **Cooperação:** O neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados, fornecendo assim a base da cooperação entre neurônios vizinhos, de acordo com a regra WTA (*Winner-Takes-All*) (KASKI; KOHONEN, 1994). Cada vez que o BMU é adaptado, seus vizinhos também são, porém, com menor intensidade. Desta interação local entre os neurônios durante o aprendizado é que emerge a organização geral do mapa.
- **Adaptação Sináptica:** Permite que os neurônios excitados aumentem seus valores individuais, em relação ao padrão de entrada, por meio de ajustes em seus pesos sinápticos. Estes ajustes são feitos para que a probabilidade de resposta do neurônio vencedor à aplicação subsequente de um padrão de entrada similar seja aumentada.

A Figura 3.2 ilustra a arquitetura de um SOM, com vetores de dados² de dimensão p na camada de entrada e um conjunto de unidades u (neurônios) arrançados na forma de um mapa em 2D na camada de saída. Cada neurônio

²Embora Tupla, vetor de características, vetor de dados e padrão de entrada sejam sinônimos na literatura, nesta tese será adotado o uso de vetor de dados.

u tem associado um vetor de pesos (ou vetor protótipo) $w_u = [w_{u1}, w_{u2}, \dots, w_{up}]$ sendo p a dimensão desse vetor, ou seja, a mesma dos vetores de dados da camada de entrada.

Figura 3.2: Arquitetura de um SOM 2D.

Seja X um conjunto de vetores de dados de entrada. Considere que um vetor de dados $v \in X$ seja escolhido e apresentado à rede. Todos os neurônios do mapa recebem o mesmo v e é calculada a distância entre o vetor de pesos w_u de cada neurônio u ao vetor de dados v , de acordo com alguma medida de proximidade. A unidade com menor distância, ou seja, o u com protótipo w mais próximo ao vetor de dados, é eleita BMU, de acordo com a Equação 3.1.

$$\|v - w_{bmu}\| = \min_u \{\|v - w_u\|\} \quad (3.1)$$

Em seguida, os vetores protótipos são atualizados. O BMU e sua vizinhança topológica são movidos na direção do vetor de dados de entrada, na camada de entrada, como se fosse um “arraste”. Desta forma, o BMU localiza o centro de uma vizinhança topológica de neurônios que cooperam entre si.

A definição desta vizinhança topológica baseia-se na evidência neurobiológica

de que há interação lateral entre um conjunto de neurônios biológicos excitados, como nas áreas sensoriais. Um neurônio que está ativo tende a excitar mais os neurônios em sua vizinhança imediata do que os neurônios mais distantes, de forma que pode-se dizer que a força de influência da vizinhança topológica ao redor de um neurônio vencedor decai suavemente com a distância lateral.

A regra para a atualização dos vetores protótipos da unidade u é:

$$w_u(t+1) = w_u(t) + \alpha(t)h_{bmu}(t)[v - w_u(t)], \quad (3.2)$$

onde t é o tempo, $\alpha(t)$ é a taxa de aprendizado e $h_{bmu}(t)$ é o *kernel* da vizinhança centrado no neurônio vencedor. O *kernel* de vizinhança pode ser uma função *Bubble* ou Gaussiana.

A função de vizinhança *Bubble*, (Equação 3.3), define um raio a partir do neurônio vencedor e o aprendizado é permitido apenas aos neurônios que estão neste raio. Outra propriedade dessa função é que todos os neurônios abarcados são movidos igualmente, independente de suas distâncias individuais ao neurônio vencedor.

$$h_{bmu} = \begin{cases} \rho(t), & \text{se } u \in N_c \\ 0, & \text{se } u \notin N_c, \end{cases} \quad (3.3)$$

onde $\rho(t)$ é uma função decrescente com o tempo e N_c é o grupo de neurônios que estão próximos o suficiente, dentro do raio, para serem atualizados.

A função de vizinhança mais comumente usada é a Gaussiana:

$$h_{bmu}(t) = \exp^{-\frac{\|r_{bmu} - r_u\|^2}{2\sigma^2(t)}}, \quad (3.4)$$

onde r_{bmu} e r_u são as posições dos neurônios BMU e u no mapa do SOM e $\sigma(t)$ é o raio da vizinhança.

Figura 3.3: Ilustração da Vizinhança Gaussiana no Mapa.

A Figura 3.3 em (a) apresenta uma típica curva gaussiana, onde a amplitude determina o nível de cooperação entre o vencedor e seus vizinhos, e em (b) pode ser vista a abrangência da função de vizinhança em um mapa 4×4 de topologia hexagonal.

Conforme aumentam o tempo de execução do algoritmo de aprendizagem e a distância entre o BMU e u , $h_{bm u} \rightarrow 0$. Da mesma forma, a taxa de aprendizado $\alpha(t)$ e o raio da vizinhança $\sigma(t)$ diminuem monotonicamente com o tempo.

Na Figura 3.4 são apresentadas a atualização do neurônio vencedor e sua vizinhança em direção ao vetor de dados na camada de entrada. Os círculos em preto e cinza correspondem às situações antes e depois da atualização, respectivamente. As linhas sólidas e pontilhadas mostram a relação da vizinhança. Por conta da relação de vizinhança e da função de vizinhança, determinados neurônios são puxados à mesma direção, de forma tal que neurônios vencedores de unidades vizinhas no mapa são semelhantes.

Figura 3.4: Atualização do neurônio vencedor e de seus vizinhos.
(VESANTO, 1997)

3.2.1 O Processo de Treinamento

Durante o treinamento, o SOM comporta-se como uma rede flexível, a qual se dobra dentro das nuvens formadas pelos vetores de dados envolvidas no treinamento. Devido à relação de vizinhança, neurônios vizinhos são “arrastados” para a mesma direção, como já dito, indicando que os vetores protótipos dos neurônios vizinhos tornam-se, em algum grau, similares (próximos espacialmente) durante o processo de aprendizado. Na interpretação de (ZUCHINI, 2003), “...SOM comporta-se como uma grade composta de neurônios ligados entre si por conexões elásticas, responsáveis por dobrar, comprimir ou esticar a grade de forma a representar, da melhor maneira possível, o conjunto de dados no espaço original”.

A cada nova apresentação de um dado ao mapa, o perfil de “sensibilidade” - ou atualização - dos neurônios vai se alterando, fato denominado processo de aprendizado ou treinamento da rede. Estas alterações são cada vez menores, de forma que a configuração do mapa convirja para uma disposição estável. O processo de aprendizado pode ser incremental (*on-line*) - (Algoritmo 2), detalhado à frente, que atualiza os pesos sinápticos quando da apresentação de um novo item de dado à rede, ou em lote (*batch*), (Algoritmo 3), detalhado adiante, que atualiza os pesos somente após cada vez que todos os dados forem apresentados à rede.

Algorithm 2 - Algoritmo Treinamento Incremental

1. Inicialização do vetor de pesos w_u do neurônio u . Faça $t = 0$, inicialize o contador de iterações ($i = 0$), $X' = X$ e inicialize $\sigma(i)$ e $\alpha(i)$.
2. Selecione, de forma aleatória, um vetor de dados v do conjunto X' e faça $X' = X' - v$.
3. Selecione o BMU de acordo com a Equação 3.1.
4. Atualize o BMU e sua vizinhança por meio da Equação 3.2.
5. Faça $t = t + 1$ e retorne ao passo 2 enquanto $X' \neq \emptyset$.
6. Incremente i . Se não atingiu o limite máximo de iterações pré-estabelecido, faça $t = 0$, $X' = X$ e ajuste $\sigma(i)$ e $\alpha(i)$. Retorne ao passo 2.

No algoritmo em lote, cada neurônio acumula as contribuições parciais de cada vetor de dados v apresentado ao mapa para os quais ele é BMU, durante uma época³ de treinamento.

Ao final de cada época os neurônios são adaptados, de acordo com uma variação da Equação 3.2, que segue:

$$w_u(t+1) = \frac{\sum_{j=1}^n h_{bmu}(t)v_j}{\sum_{j=1}^n h_{bmu}(t)}, \quad (3.5)$$

onde BMU é o neurônio vencedor para o vetor de dados v_j , conforme (3.1), e o peso de cada dado é estabelecido segundo o valor da função de vizinhança $h_{bmu}(t)$ para cada neurônio vencedor.

Algorithm 3 - Algoritmo Treinamento em Lote

1. Inicialização do vetor de pesos w_u do neurônio u . Faça $t = 0$, inicialize o número de iterações ($i = 0$), $X' = X$ e inicializa $\sigma(i)$. A taxa de aprendizado $\alpha(i)$ deve receber um valor pequeno e fixo.
2. Selecione um vetor de dados do conjunto X' e faça $X' = X' - v$.
3. Selecione o BMU de acordo com a Equação 3.1.
4. Calcule a contribuição parcial do vetor de dados para o neurônio BMU e sua vizinhança.
5. Retorne ao passo 2 enquanto $X' \neq \emptyset$.
6. Os neurônios e seus vizinhos são atualizados (3.5).
7. Incremente i . Se não atingiu o limite máximo de iterações pré-estabelecido, faça $X' = X$ e ajuste $\sigma(i)$ e $\alpha(i)$. Retorne ao passo 2.

³Uma época corresponde à apresentação de todos os padrões de treinamento disponíveis à RNA.

Ambos os algoritmos de treinamento operam com duas fases bem distintas, de ordenação inicial (fase mais rápida) e de convergência (de processamento mais demorado), para as quais os valores de α e σ devem ser grandes na primeira fase e pequenos na segunda fase, para o alcance de um ajuste mais refinado do mapa. Na fase de ordenação, os neurônios se posicionam no mapa topológico e na fase de convergência a acuidade estatística na localização dos neurônios é feita, possibilitando uma melhor representação dos dados.

A Figura 3.5, gerada na SOM-Toolbox[©], ilustra um treinamento sequencial em trezentas épocas que mapeia os dados sobre os neurônios, tal que o ajuste elástico e gradual da rede é destacado. As cruces em vermelho são os vetores dados de entrada. Os pontos pretos mostram as posições das unidades do mapa e as linhas cinza mostram as conexões de vizinhança dessas unidades. Inicialmente, as posições dos neurônios no espaço de entrada podem estar completamente desordenadas e distantes dos dados a serem mapeados. Durante o treinamento o mapa se organiza, e se adequa aos dados de treinamento.

A despeito de uma fase de convergência bem definida no treinamento, não há, porém, uma função de erro geral (ou de energia) a ser minimizada em SOM, de forma que o estado estacionário (de convergência) do mapa possa ser garantido. Há apenas provas de convergência para um arranjo unidimensional (KOHONEN, 1982), (RITTER; SCHULTEN, 1986) e uma tentativa de generalização do processo por Cadeias de *Markov* em (RITTER; SCHULTEN, 1988). Erwin e colaboradores demonstraram em seus estudos (ERWIN; OBERMAYER; SCHULTEN, 1992a) e (ERWIN; OBERMAYER; SCHULTEN, 1992b)⁴ que tais funções não podem existir para casos genéricos, impondo um obstáculo à análise matemática do processo de convergência do SOM.

⁴Para uma discussão aprofundada sobre convergência do SOM, essas são referências indicadas.

Figura 3.5: Exemplo de um mapa sendo treinado para dados e protótipos bidimensionais. O mapa está inicialmente ajustado em uma região não superposta à dos dados, para evidenciar a movimentação dos seus protótipos para se ajustar aos dados, interativamente.

3.2.2 Questões sobre Parametrização do Mapa

Como outros algoritmos de aprendizagem não-supervisionada, SOM apresenta vários parâmetros a serem definidos para sua boa execução. Definir os parâmetros de treinamento corretos não é uma tarefa trivial e não há critérios bem definidos para a escolha de parâmetros que produzam mapas bem ajustados mas sim, heurísticas propostas para este fim. Há portanto, um consenso entre os pesquisadores de se testar diferentes configurações de mapas e só então, decidir pela melhor solução, segundo alguma medida de qualidade do mapa ou de validação, no caso de análise de agrupamentos.

Abaixo, as principais estratégias encontradas na literatura sobre a parametrização do SOM são apresentadas, de forma a ajudar o leitor na sua definição.

- **Encontrar a dimensão ideal para o mapa.** Mapas bidimensionais são os de utilização geral pela comunidade acadêmica para análise exploratória, pela sua facilidade de visualização e interpretação. Estudos como o de (PERES, 2006) advogam que para situações particulares, dimensões maiores podem ser indicadas. Definir as dimensões do mapa a partir de uma análise visual prévia sobre os dados também é uma estratégia, segundo (KOHONEN, 2001).
- **Encontrar o número ideal de neurônios.** O número de neurônios determina a granularidade do resultado do mapeamento, que afeta o custo computacional, a qualidade e a capacidade de generalização do SOM. A determinação do número de neurônios é empírica, de forma que alguns autores tentam estabelecer heurísticas para tal fim. Por exemplo, (VESANTO et al., 2000) afirma, a partir de seus experimentos que, sendo n o tamanho do conjunto de dados, a fórmula $q = 5\sqrt{n}$, provê um bom número de neurônios para o mapa. Para (KASKI, 1997), se o número de vetores de dados de entrada for pequeno (< 1000), pode-se adotar o número de neurônios igual ao dos dados.
- **Definir o formato do mapa.** O mapa pode assumir formato retangular, hexagonal, cilíndrico e toroidal. Uma vez que visualização de dados é a principal aplicação de SOM, arranjos hexagonais favorecem a inspeção visual, justifica (KOHONEN, 2001).
- **Definir função de vizinhança.** O *Kernel* de vizinhança gaussiano é o mais indicado, pelo fato de, normalmente, evidenciar melhor tendência de agrupamentos em ferramentas visuais de análise;
- **Inicializar vetores de pesos (protótipos).** A inicialização pode ser feita de três formas: *aleatória*, na qual os vetores de pesos são inicializados

com pequenos valores randômicos; *com exemplos*, onde os vetores de pesos são inicializados com exemplos aleatoriamente sorteados do conjunto de dados de entrada; ou *linear*, que consiste em determinar os dois autovetores da matriz de autocorrelação do conjunto de dados que possuem os maiores autovalores (componentes principais), e então expandir um subespaço linear de duas dimensões sobre os neurônios do mapa. Efetuar inicialização linear é a sugestão de (KOHONEN, 2001), uma vez que se tem a intenção de impor alguma ordem já na criação do mapa. Assim, o mapa que já começa de certa forma organizado, e consegue atingir uma configuração estável em muito menos iterações que em outras formas de inicialização.

- **Escolher algoritmo de treinamento.** O treinamento seqüencial (*on-line*) é bastante utilizado, por compor o SOM original. Utilizar algoritmo de treinamento em lote, contudo, também traz bons resultados. Pode-se comparar os resultados de ambos os treinamentos para a escolha final. O mais importante, porém, é escolher taxas de treinamento adequadas. Usar $\alpha(t) = 0,5$ na primeira fase e $\alpha(t) = 0,05$ na fase de convergência, é um bom exemplo de taxa de aprendizado adequada (KASKI, 1997), ou seja, taxas maiores na primeira fase e menores na segunda, para um ajuste mais fino.
- **Determinar o número de épocas.** Usar como número de épocas 10, 100, 1.000 e 10.000 é uma boa prática. Estimar o número de épocas de treinamento n_e é uma proposta de (VESANTO et al., 2000), tal que $n_e = 10 \frac{q}{n}$ para a fase inicial e $n_e = 40 \frac{q}{n}$ para a fase de convergência, onde q é o número de neurônios do mapa e n o tamanho do conjunto de dados de entrada. Essa proposição é empírica.

3.3 Propriedades de um SOM

De acordo com (HAYKIN, 1999), (VESANTO, 1997) e (KOHONEN, 2001), o mapa gerado pelo algoritmo SOM tem as seguintes propriedades importantes:

- **Aproximação dos Dados de Entrada:** O mapa, representado pelo conjunto de vetores de pesos sinápticos no espaço de saída, fornece uma boa aproximação aos dados de entrada. Isso se dá por meio da quantização do espaço.
- **Ordenação Topológica:** É ordenado topologicamente, no sentido de que a localização espacial de um neurônio no mapa corresponde a um domínio particular ou a características dos vetores de dados de entrada.
- **Casamento de Densidade:** O mapa reflete variações na estatística da distribuição da entrada. Regiões no espaço de entrada de onde os vetores de dados são presentes com uma alta probabilidade de ocorrência são mapeadas para domínios maiores do mapa de saída, e portanto, com melhor resolução que regiões com baixa probabilidade de ocorrência. A tarefa de identificação de agrupamentos em um mapa de características está diretamente relacionada a esta característica.
- **Seleção de Características:** Com os dados de entrada sob uma distribuição não-linear, o mapa auto-organizável é capaz de selecionar um conjunto das melhores características (atributos) para aproximar a distribuição subjacente, como em um processo de Análise de Componentes Principais (PCA) não-linear. A motivação para o uso de dados com dimensões reduzidas é que a complexidade computacional do passo subsequente é reduzida, por exemplo, no agrupamento.

3.4 Medidas Avaliativas para um SOM

A auto-organização do mapeamento é definida na fase de treinamento. Durante essa etapa, como já dito, várias decisões sobre os vários parâmetros do algoritmo devem ser tomadas. Esses parâmetros influenciam no mapa final, devendo ser escolhidos com cautela. Na prática, diferentes possibilidades são testadas, medindo-se a qualidade desses mapas, para então eleger o melhor, por algum critério estabelecido.

A escolha do “melhor” mapa deve ser por aquele que melhor represente os dados de entrada, ou seja, que mais satisfaça as propriedades citadas na Seção 3.3, principalmente no tocante à sua resolução e em quão bem o mapa preserva a topologia dos vetores de dados.

A avaliação topológica caracteriza uma das mais eficientes formas de mensurar a qualidade de um mapeamento realizado por um SOM. Como os dados de entrada podem não residir em um espaço de dimensão igual ao escolhido para o mapa de saída, um trabalho de redução de dimensão acontece, justificando a avaliação do mapeamento no âmbito da preservação topológica. Desta forma, para (PERES, 2006), não existe preservação de topologia perfeita, mas sim aproximações boas e ruins.

3.4.1 Quantização Vetorial

Quantizar um espaço significa dividi-lo em regiões distintas e para cada uma destas definir um vetor protótipo (*codebook*, ou vetor de reconstrução). Novos pontos deste espaço de quantização poderão ser identificados, de acordo com sua proximidade ao vetor protótipo de cada região.

Quantização Vetorial é um método clássico de aproximação de sinais que geralmente forma uma aproximação quantizada da distribuição dos vetores de

dados de entrada $x \in \mathfrak{R}$, usando um número finito de vetores protótipos $w_u \in \mathfrak{R}$, $u = 1, 2, \dots, n$. A aproximação de x significa encontrar o vetor protótipo w_u mais próximo a x no espaço de entrada. Normalmente, usa-se para isso a distância euclidiana (KOHONEN, 2001).

A quantização de um SOM é mensurada pelo **Erro de Quantização** (E_q), que está baseado no cálculo das médias das distâncias entre os vetores de dados e seus *BMUs*. Esse erro correspondente à diferença (distorção) entre o vetor de dados v e o vetor de pesos w_{bmu} , vencedor no processo de competição pelo padrão v , e é dada pela Equação 3.6.

$$E_q = \frac{\sum_{k=1}^n \|v_k - w_{bmu}\|}{n} \quad (3.6)$$

onde n representa o tamanho do conjunto de dados.

O Erro de Quantização é uma medida de acuidade ou resolução do mapa inversamente proporcional ao número de neurônios, ou seja, o erro diminui com o aumento de neurônios no mapa. O menor erro indica maior proximidade dos dados de seus neurônios vencedores, ou seja, estão melhor representados. Contudo, se o mapa possuir um número muito grande de neurônios ou sofrer um processo de treinamento onde o raio de vizinhança tornou-se menor ou igual a 1 por muito tempo, pode acontecer dos neurônios posicionarem-se sobre (ou praticamente sobre) os dados a serem representados, de forma que $E_q \rightarrow 0$ mas com o arranjo tão retorcido que a capacidade de representar a topologia dos dados é perdida.

3.4.2 Topologia

Topologia, no contexto de análise de SOM, é uma forma de expressar as similaridades existentes entre as características do vetor de dados mapeado, por meio de pontos localizados proximamente em um mapa neural. É a preservação da

vizinhança na projeção (URIARTE; MARTÍN, 2006). De forma coloquial, a topologia de um mapa significa o mapeamento de dados semelhantes para posições próximas no mapa.

A topologia pode ser vista como a caracterização do inter-relacionamento entre os dados. Sua preservação é bastante importante, e, de acordo com (BAUER; HERRMANN; VILLMANN, 1999), pode ser explorada para reduzir as consequências de ruído em uma transmissão de dados quantizados, para melhorar a visualização de dados de alta dimensionalidade, ou em muitas outras aplicações que fazem uso da distância entre pontos de dados.

De acordo com (GOODHILL; SEJNOWSKI, 1996), uma forma de manutenção de relações de similaridade está em nível de ordenação dessas relações, de forma que a comparação entre os valores das similaridades entre os pontos no espaço de entrada com os valores das similaridades entre os pontos no espaço de saída deve verificar a mesma relação de ordenação. As medidas de avaliação de topologia descritas a seguir estão ligadas a esse quesito.

O **Erro Topológico** (E_t) avalia o quanto a estrutura do mapa aproxima vetores de dados de entrada próximos, a partir do primeiro e do segundo neurônios BMUs, de forma que corresponderá ao percentual de vetores de dados cujos primeiro e segundo BMUs não são vizinhos no mapa.

$$E_t = \frac{1}{n} \sum_{k=1}^n u(v_k) \quad (3.7)$$

$$\text{em que } u(v_k) = \begin{cases} 1, & \text{se o primeiro e segundo BMUs não forem adjacentes;} \\ 0, & \text{caso contrário.} \end{cases}$$

Portanto, o erro topológico indica a dificuldade do mapa em representar relações de vizinhança entre os dados, e por isso, indica uma necessidade de análise ou melhoria no mapeamento, mas não indica, necessariamente, que a quantização

esteja ruim.

Outra forma de averiguar a preservação da vizinhança é computar o **Produto Topológico** (P_t), cujo objetivo é quantizar a preservação de relações de vizinhança em mapeamentos realizados sobre espaços de dimensões iguais ou não. Esse processo de medição de qualidade da topologia relaciona, para cada neurônio, a seqüência de vizinhos no espaço de entrada com a seqüência de vizinhos no espaço de saída (BAUER; PAWELZIK, 1992).

O Produto Topológico é formalmente descrito em (3.8)⁵. Na formulação, d_V representa as relações de vizinhança dos vetores de dados de entrada e d_A a vizinhança no mapa de saída.

$$P_t = \frac{1}{N^2 - n} \sum_{j=1}^N \sum_{k=1}^{N-1} \log \left(\prod_{l=1}^k \frac{d_V(w_j, w_{n_l^A(j)})}{d_V(w_j, w_{n_l^V(j)})} \frac{d_A(j, n_l^A(j))}{d_A(j, n_l^V(j))} \right)^{\frac{1}{2k}} \quad (3.8)$$

As relações de vizinhança $n_l^V(j)$ e $n_l^A(j)$ representam, respectivamente, o l -ésimo vizinho mais próximo do neurônio j no espaço de entrada do mapeamento e o l -ésimo vizinho mais próximo do neurônio j no espaço de saída do mapeamento. Essas relações são dadas pelo cálculo dos vizinhos mais próximos.

Os termos do produtório da Equação (3.8) são formados pelo cálculo das distâncias entre os neurônios e seus vizinhos, como segue:

1. $d_V(w_j, w_{n_l^A(j)})$: Calcula a distância entre um neurônio j e seu l -ésimo vizinho mais próximo no espaço de saída, ambos representados por seus vetores de pesos. Mede-se no espaço de entrada do mapeamento a distância existente entre um neurônio e seu vizinho, buscando distorções de topologia.

⁵O Produto Topológico foi primeiramente introduzido no contexto de dinâmica não-linear e análise de séries temporais, com o nome de *Wavering Product*. Para uma discussão mais detalhada sobre essa medida (BAUER; PAWELZIK, 1992) e (BAUER; HERRMANN; VILLMANN, 1999), utilizados nesta seção, devem ser consultados.

2. $d_V(w_j, w_{n_l^V(j)})$: Calcula a distância entre um neurônio j e seu l -ésimo vizinho mais próximo no espaço de entrada, ambos representados por seus vetores de pesos.
3. $d_A(j, n_l^A(j))$: Calcula a distância entre um neurônio j e seu l -ésimo vizinho mais próximo no espaço de saída, ambos representados pelo índice do neurônio na matriz que determina a vizinhança topológica.
4. $d_A(j, n_l^V(j))$: Calcula a distância entre um neurônio j e seu l -ésimo vizinho mais próximo no espaço de entrada, ambos representados pelo índice do neurônio na matriz que determina as relações de vizinhança topológica. A partir desse cálculo, busca-se avaliar o quanto o vizinho no espaço de saída está distante no espaço de entrada, procurando por violações de topologia.

Utilizando essa indexação de vizinhança mais próxima, as seguintes relações podem ser estabelecidas:

$$R_V(j, l) = \frac{d_V(w_j, w_{n_l^A(j)})}{d_V(w_j, w_{n_l^V(j)})} \quad (3.9)$$

$$R_A(j, l) = \frac{d_A(j, n_l^A(j))}{d_A(j, n_l^V(j))} \quad (3.10)$$

Desta definição tem-se que $R_V(j, l) = R_A(j, l) = 1$ somente se os vizinhos mais próximos de ordem l nos espaços de entrada e saída coincidirem, ou seja, $n_l^V(j) = n_l^A(j)$ e $w_{n_l^V(j)} = w_{n_l^A(j)}$. Qualquer desvio de R_V e R_A do valor 1 indicará violação da ordenação de vizinhança feita no mapa, produzindo contração ou contorsão no mesmo.

Os termos até então discutidos são calculados para todos os neurônios, e os valores obtidos estão relacionados em um produtório. A razão para este tipo de relação é a necessidade de anular efeitos de distorções topológicas muito peque-

nas, com efeito local ((BAUER; HERRMANN; VILLMANN, 1999)). Uma medida de distorções topológicas pode ser obtida da combinação das relações de forma multiplicativa, proporcionando uma espécie de ponderação entre as medidas obtidas nos dois ângulos de observação:

$$R(j, k) = \left(\prod_{l=1}^k R_V(j, l) R_A(j, l) \right)^{\frac{1}{2k}} \quad (3.11)$$

Conforme (BAUER; HERRMANN; VILLMANN, 1999), chega-se assim ao tipo de distorção topológica existente. Valores > 1 para $R(j, k)$ são obtidos quando as maiores distâncias são observadas no espaço de saída, ou seja, o espaço de saída se contrai em um fator maior que o espaço de entrada, mostrando que sua dimensão é mais alta. O caso inverso, quando as distâncias observadas no espaço de saída são menores que no espaço de entrada, levam a um valor < 1 e mostram que o espaço de saída está contraído e possui uma dimensão menor.

Finalmente, para análise de todo o mapeamento, os raios descritos são calculados para todos os neurônios do mapa e em relação a todas as suas ordens de vizinhança. Os somatórios executam esse procedimento, sendo que o primeiro (em j) diz respeito aos neurônios e o segundo (em k), aos seus vizinhos. A função logaritmo é utilizada para facilitar as análises de desvios de 1 e uma normalização em relação ao número de neurônios na rede é aplicada.

Outras medidas podem ser encontradas para medir a qualidade de topologia de um mapeamento, como em (VILLMANN et al., 1997), (BAUER; HERRMANN; VILLMANN, 1999), (AROUS; ELLOUZE, 2003) e (PÖLZLBAUER, 2004).

3.5 Considerações Finais

Por possuir um robusto conjunto de protótipos, que de forma aproximada segue a função de densidade de probabilidade dos dados de entrada, e ter a independência

da ordem dos vetores de dados de entrada como méritos, SOM pode ser usado em agrupamento e análise de dados, a exemplo dos trabalhos de (KASKI, 1997), (ALHONIEMI, 2002), (PARENTE et al., 2003), (BOSCARIOLI; SILVA; HERNANDEZ, 2006a) e (BOSCARIOLI; SILVA; HERNANDEZ, 2006b).

No entanto, SOM possui também um número de parâmetros dependentes do usuário que podem causar problemas quando aplicados em situações práticas reais e, por vezes, torna-se necessário iterar vários treinamentos de SOM, com diferentes parâmetros e validá-los com diferentes tipos de medidas de qualidade.

Embora neste trabalho enfatiza-se o SOM tradicional, há muitas variações para esse algoritmo. O próprio inventor, *Teuvo Kohonen*, em (KOHONEN, 2001) faz um apanhado dessas variantes. Segundo (VESANTO, 2000), todas essas variações são, essencialmente, protótipos e um conjunto de relações de vizinhança definidas entre eles. Os vetores protótipos são interativamente ajustados para corresponder aos dados de treinamento, e as relações de vizinhança são usadas de forma que os vetores protótipos vizinhos sejam semelhantes.

SOM é o modelo de RNA competitiva mais utilizado, nas várias áreas do conhecimento, conforme atestam (KASKI; KANGAS; KOHONEN, 1998), (OJA; KASKI; KOHONEN, 2002) e (PÖLLÄ, M.; HONKELA; KOHONEN, 2007). O modelo de *Kohonen* pode ser usado nas diversas aplicações em que comumente se utilizam outros modelos de redes neurais, como reconhecimento de padrões e processamento de sinais, sendo eficiente em aplicações onde os padrões possuam alguma relação entre si, podendo desta forma, serem classificados (KOHONEN, 2001).

Este capítulo procurou descrever SOM em seus aspectos principais, enfatizando questões importantes de projeto, como sua parametrização e a avaliação da qualidade do mapeamento.

4 Análise Exploratória via Mapa Auto-Organizável

“Você tem o seu caminho. Eu tenho o meu caminho. Quanto ao caminho exato, o caminho correto, e o único caminho, isso não existe.” (*Friedrich Nietzsche*)

Este capítulo tem como objetivos abordar o uso de SOM em análise exploratória com ênfase na tarefa de análise de agrupamentos, apresentar trabalhos correlatos a esta pesquisa e descrever conceitualmente a metodologia de análise de agrupamentos idealizada nesta pesquisa.

4.1 Conceitos Gerais

A análise exploratória de dados é uma abordagem para análise de dados que emprega uma variedade de técnicas de visualização e análise de agrupamentos na busca por padrões em uma base de dados. A Figura 4.1 esquematiza o uso de SOM na análise de dados, onde, primeiro se faz a redução do conjunto de dados para vetores protótipos, os quais serão utilizados, a posteriori, por outras técnicas, sejam de visualização ou de agrupamento dos dados.

Deve-se deixar claro que, com exceção de abordagens como as citadas nas

Figura 4.1: Análise de Dados a partir do SOM.

próximas seções, a saída de um mapa SOM fornece apenas a representação dos dados via organização topológica dos neurônios/protótipos. Este resultado nem sempre é passível de uma análise visual direta, sendo, portanto, difícil sugerir a estrutura de grupos existente, haja vista o mapeamento topológico do SOM não ser suficiente para realizar uma análise de agrupamentos, o que remete à aplicação de outras técnicas sobre o SOM para que os resultados (grupos) sejam, de alguma forma, observáveis e úteis na geração de novos conhecimentos.

A saída de um SOM para um vetor de dados inserido é geralmente o índice do *BMU*, no caso bidimensional um par de valores (i, j) , e o nível de ativação diretamente relacionado à quantização (a distância do padrão de entrada ao neurônio vencedor, computada no espaço de pesos), (COSTA, 1999).

A Figura 4.2 mostra os vetores de peso (protótipos) no espaço de dados de entrada, após o treinamento do SOM para um exemplo ilustrativo com dados bidimensionais. Ao observar a densidade dos neurônios da rede, pode-se sugerir a existência de três grupos, circutados em vermelho, mas essa asserção nem sempre é intuitiva como nesse exemplo.

Figura 4.2: Mapa topológico gerado por SOM. As coordenadas x_1 e x_2 representam as dimensões dos vetores de dados de entrada. Cada ponto representa um protótipo no mapa.

Para se ter informações sobre agrupamentos é necessário um processamento sobre os neurônios *BMUs*. Para tal, existem diversas abordagens, destacando-se a construção da Matriz-U (técnica descrita na próxima Seção) e a aplicação de técnicas de agrupamento sobre os neurônios.

Pela ótica da segunda abordagem, pode-se apresentar os dados ao mapa treinado, de forma a classificá-los, desde que se tenham rótulos para os dados, uma vez que de posse do mapa rotulado ou analisado, consegue-se a informação do grupo a que um dado pertence.

Por exemplo, considere um conjunto de dados sintéticos, com a distribuição mostrada na Figura 4.3. Considere ainda que A, B e C sejam, respectivamente, rótulos dos três grupos existentes. Ao treinar um mapa SOM 5×5 para esses dados, pode-se efetuar a rotulação e a contagem da frequência dos dados representados por cada neurônio, e obter a abstração do mapa da Figura 4.4. Esta é uma aplicação típica para um mapa SOM.

Neste exemplo, onde há rótulos definidos previamente, é factível a interpre-

tação da existência de grupos. No entanto, sem a existência desses rótulos, ou mesmo quando os grupos não são linearmente separáveis, não é possível afirmar que neurônios, muito embora vizinhos, representam dados que estejam contidos em um mesmo grupo, nem sequer a quantidade de grupos presentes.

Figura 4.3: Distribuição dos dados de entrada.

Além da redução do conjunto de dados para análise, outra vantagem de SOM é que não é necessário recomputar o mapa para cada novo dado de entrada, pois, se as estatísticas podem ser assumidas como estacionárias, um novo dado pode ser diretamente mapeado ao vetor protótipo representante do item de dados mais próximo no modelo antigo. Ou seja, uma vez de posse de um mapa treinado e dos grupos detectados, SOM pode ser usado para classificação de um novo vetor de dados.

Figura 4.4: Mapa SOM rotulado, onde letras indicam rótulos e os números entre parênteses a contagem dos vetores de dados representados por cada protótipo.

4.2 Visualização de um Mapa Auto-Organizável

Como já abordado no capítulo anterior, um dos objetivos de SOM é representar padrões de entrada de alta dimensionalidade com vetores protótipos, para que possam ser visualizados, de forma facilitada, em um mapa de dimensão menor, geralmente $d_i = 2$, uma vez que o número limitado de dimensões visuais é um dos problemas da visualização de dados multidimensionais.

A visualização dos dados é, em muitos casos, uma informação valiosa, inclusive quando se deseja descobrir a quantidade de grupos em uma base de dados.

Como uma ferramenta de auxílio à análise visual, uma técnica conhecida por Matriz-U (*U-Matrix*) foi proposta em (ULTSCH, 1992) e (ULTSCH, 1993) para, a partir de um mapa treinado, detectar visualmente informações intrínsecas dos vetores protótipos, como a relação topológica e o número de grupos.

A idéia central é medir as distâncias entre pesos sinápticos de neurônios adjacentes para construir a Matriz-U, e interpretá-la como uma imagem $f(x, y)$. Nesta imagem, as coordenadas de cada *pixel* (x, y) são derivadas das coordenadas dos neurônios do mapa (por exemplo, $(1, 1), (1, 2) \dots (X, Y) \rightarrow (1, 1), (1, 2) \dots (2 * X - 1, 2 * Y - 1)$), na qual a intensidade de cada *pixel* corresponde à distância calculada.

Para entender a construção da Matriz-U, suponha um mapa cujas unidades possuem vizinhança retangular, e cujo tamanho seja $X \times Y$. Seja $[b_{x,y}]$ a unidade em avaliação no mapa. Para cada unidade existem três distâncias a seus vizinhos $(d_x, d_y$ e $d_{xy})$, que devem ser armazenadas na Matriz-U, como visto na Figura 4.5. Observe que esse raciocínio pode ser estendido para construção da Matriz-U para um mapa de vizinhança hexagonal.

Considerando que se está utilizando a distância Euclidiana, o cálculo de d_x , d_y e d_{xy} é como nas Equações 4.1, 4.2 e 4.3.

Figura 4.5: Exemplo de construção da Matriz-U para uma estrutura de vizinhança retangular, onde se vê as três distâncias calculadas.

Adaptada de (COSTA, 1999).

$$dx(x, y) = \|b_{x,y} - b_{x+1,y}\| \quad (4.1)$$

$$dy(x, y) = \|b_{x,y} - b_{x,y+1}\| \quad (4.2)$$

$$dxy(x, y) = \frac{1}{2} \left(\frac{\|b_{x,y} - b_{x+1,y+1}\|}{\sqrt{2}} + \frac{\|b_{x,y+1} - b_{x+1,y}\|}{\sqrt{2}} \right) \quad (4.3)$$

As distâncias calculadas são armazenadas na Matriz-U e plotadas em uma estrutura matricial de tamanho $(2*X-1) \times (2*Y-1)$, de acordo com o esquema mostrado na Figura 4.6. Para maiores informações sobre Matriz-U, recomenda-se (ULTSCH; SIEMON, 1990), (ULTSCH, 1992), (ULTSCH, 1993) e (COSTA, 1999).

Figura 4.6: Esquema de preenchimento e plotagem de uma Matriz-U.
Adaptada de (COSTA, 1999).

Em uma Matriz-U bidimensional, pode-se visualizar agrupamentos e regiões de separação. Na Figura 4.7 (a), observa-se, claramente, três regiões - partes mais claras - com bordas bastante detectáveis, as regiões mais escuras. É fato que pode haver grupos menos separados, o que tornará bastante difícil a interpretação da Matriz-U.

Outra forma de plotar a Matriz-U é em um gráfico 3D, conforme Figura 4.7 (b), em que o valor do *pixel* na coordenada (x, y) é representado por um ponto na coordenada z , onde pode-se verificar picos montanhosos e vales. Topologicamente, as distâncias entre neurônios vizinhos refletem os agrupamentos, pois

um vale da superfície de relevo representa neurônios pertencentes a um mesmo grupo. Neurônios que têm uma grande distância em relação ao seu adjacente são representados por picos da superfície de relevo, que podem ser associados a regiões ou unidades em fronteiras de agrupamento.

(a)

(b)

Figura 4.7: Exemplos de Matriz-U: (a) Matriz bidimensional em tons de cinza; (b) Matriz visualizada com representação por superfície topológica.

Outras técnicas podem ser utilizadas na visualização de um mapa auto-organizado. Se os mapas tiverem dimensão $d_i \leq 3$, é possível usá-los como co-

ordenadas no espaço $\mathbb{R}^{di}$ para visualizar a disposição dos neurônios. (VESANTO, 1999) sugere Mapa de *Sammon* e outros métodos de projeção para esse fim. Outras técnicas de visualização para dados com alta dimensionalidade são abordadas em (GRINSTEIN; TRUTSCHL; CVEK, 2001) e (OLIVEIRA; LEVKOWITZ, 2003).

4.3 Agrupamento de dados por Mapa Auto-Organizável

O algoritmo SOM tem sido, há alguns anos, utilizado como base para o desenvolvimento de várias metodologias para análise de agrupamento de dados. Em (LAMPINEN; OJA, 1992), um SOM em duas camadas foi proposto, para agrupar dados. A segunda camada do SOM toma como entrada os protótipos, organizados por similaridade, na primeira camada. Nesta abordagem, o número de neurônios no segundo mapa é igual ao número de grupos desejados, de forma que a tarefa dessa segunda camada é análoga ao algoritmo *K-means*. Há trabalhos como o de (BAÇÃO; LOBO; PAINHO, 2005) que propõem a utilização de SOM unidimensional como substituto ao *k-means*, assumindo vizinhança zero ao final do aprendizado, e que cada unidade do mapa auto-organizado é o centro de um grupo.

Alguns trabalhos que especificamente ocupam-se da tarefa de agrupamento de dados via SOM sem conhecimento prévio do número de grupos desejados, e considerados correlatos a esta pesquisa, seguem descritos, bem como a metodologia de análise de agrupamento aqui proposta.

4.3.1 SL-SOM *Self-Labeled SOM*

Neste método, proposto por (COSTA, 1999), a idéia é, a partir de um mapa SOM treinado particionar os *pixels* da Matriz-U usando o algoritmo *Watershed*, uma ferramenta de segmentação de imagens utilizada dentro da morfologia

matemática, proposta por (BEUCHER; LANTUÉJOUL, 1979).

A idéia básica de *Watershed*, ou linha de partição de águas, pode ser exemplificada por uma analogia: Uma imagem em níveis de cinza pode ser vista como uma superfície cujos pontos têm altitudes relativas aos valores dos *pixels*. Quanto maior o valor do nível de cinza correspondente, maior a altitude. Como as bordas dos objetos são previamente ressaltadas, a superfície será formada de vales e bacias. Ainda, pode-se imaginar que essa superfície é imersa em água, a partir dos pontos mais baixos. Conforme o nível de água nas bacias e vales aumenta, as águas dessas bacias transbordam e tocam aquelas provenientes de outras bacias. Nesse ponto, são construídas barragens para evitar a união dessas águas, daí o nome de “linha de partição de águas”. O algoritmo marca os *pixels* que têm possibilidade de receber dois rótulos diferentes, o que indica que as águas de duas bacias distintas se encontraram (COSTA, 1999).

Há os seguintes passos a serem seguidos:

1. Obter a Matriz-U.
2. Encontrar os marcadores para a Matriz-U, por meio da detecção de regiões conectadas.
3. Aplicar o algoritmo *Watershed* sobre a Matriz-U, utilizando-se dos marcadores obtidos no passo 2.
4. Rotular as regiões conectadas da imagem segmentada no passo 2.
5. Copiar os rótulos obtidos no passo 4 aos neurônios associados a cada *pixel* da Matriz-U.
6. Rotular possíveis neurônios ainda não rotulados, pelo vizinho mais próximo ou pelo *K-means*.

O grande diferencial deste algoritmo é utilizar *Watershed*, uma técnica de segmentação de imagens, para particionar a Matriz-U em regiões conectadas. Como limitações, pode-se citar que sua aplicação não é recomendada para mapas com poucos neurônios, ou para casos onde os agrupamentos estejam menos separados. A baixa separabilidade causa uma degradação das bordas da Matriz-U, o que pode tornar sua segmentação um processo bastante difícil.

4.3.2 Metodologia de Vesanto & Alhoniemi

A proposta de (VESANTO; ALHONIEMI, 2000) é a descoberta de agrupamentos em duas fases: Treinar uma rede SOM a partir dos dados e então utilizar os vetores de pesos dessa rede para descobrir agrupamentos, a partir do algoritmo *K-means*.

Primeiramente, um grande conjunto de protótipos, muito maior que o número de grupos esperado é formado pelo SOM. Esses protótipos são então combinados para formar grupos. Os neurônios que não representam nenhum elemento do conjunto de dados, denominados unidades de interpolação, são excluídos, como uma espécie de pré-processamento. Como benefício principal dessa proposta, os autores apontam a redução do custo computacional ao algoritmo de agrupamento (2ª fase). Em outras palavras, a aplicação de SOM funciona como uma espécie de pré-processamento para o agrupamento.

Dada a sensibilidade na inicialização do algoritmo *k-means*, executa-se 100 vezes esse algoritmo para cada valor de k , com inicialização randômica. O melhor particionamento para cada número de grupos é selecionado pelo critério de erro abaixo:

$$E = \sum_{k=1}^C \sum_{x \in Q_k} \{\|x - c_k\|\}^2, \quad (4.4)$$

onde C é o número de grupos e c_k o centro de cada grupo.

A partir daí, o índice *Davies-Bouldin* (ver Anexo) é aplicado para selecionar o

melhor agrupamento entre os particionamentos com diferentes números de grupos.

Um grande problema dessa abordagem é que ignora grupos com apenas um elemento (*singleton clusters*), que podem ser vistos como grupos raros. Outro problema, é a utilização do *K-means* e/ou do índice *Davies-Bouldin*, que calculam o grau de separação entre grupos por meio de centróides, impondo uma geometria esférica aos grupos, o que nem sempre se encontra em dados reais.

Os trabalhos de (ALHONIEMI, 2002), (ZUCHINI, 2003), (HIMBERG, 2004), (PERES, 2006) e (SASSI, 2006) - entre vários outros trabalhos de análise exploratória com SOM - seguem esta metodologia em seus experimentos.

Pode-se considerar essa metodologia, a despeito de seus limitantes, uma das mais utilizadas para agrupamento de dados por meio de SOM, haja vista ela estar implementada na *SOM-Toolbox* (VESANTO et al., 2000), uma ferramenta proposta e divulgada pelo Centro de Pesquisas em Redes Neurais da Universidade de Tecnologias de Helsinki/Finlândia, que é amplamente utilizada pela comunidade acadêmica para fins de ensino e pesquisas, seja para geração de agrupamentos ou para *Benchmarking*.

4.3.3 Agrupamento baseado em Particionamento de Grafos

Este algoritmo, proposto por (COSTA; NETTO, 2003) baseia-se nas informações de distância entre os neurônios e no nível de atividade do neurônio¹. Tem como estratégia inicial considerar que todos os neurônios fazem parte de um grafo não orientado, totalmente conectado. Com base em heurísticas, arestas consideradas conexões inconsistentes entre neurônios vizinhos são eliminadas, tendo como produto final vários grafos menores, que representam agrupamentos distintos.

¹O nível de atividade se refere ao número de vezes que um neurônio foi *BMU*, o que nada mais é que um histograma de "disparos".

Os passos desse algoritmo, a partir de um mapa auto-organizado treinado, são:

1. Obter as distâncias entre os vetores de pesos dos neurônios adjacentes u e j , $d(w_u, w_j)$ e obter o nível de atividade de cada neurônio u , denominada $H(u)$. Excluir neurônios que não representam nenhum dado de entrada.
2. Para cada par de neurônios adjacentes u e j , a aresta será considerada inconsistente:
 - Se a distância entre os pesos exceder em 2 a distância média dos demais neurônios adjacentes a u ou a j ;
 - Se os dois neurônios adjacentes u e j possuírem atividade H abaixo de 50% da atividade mínima permitida H_{min} , ou um dos neurônios for inativo ($H(t) = 0$);
 - Se a distância entre os centróides dos conjuntos de dados de entrada associados aos neurônios u e j exceder em 2 vezes a distância entre os vetores de pesos $d(w_u, w_j)$.
3. Remover arestas inconsistentes. Isso corresponde a colocar o valor 0 na coordenada (u, j) da matriz de adjacência que representa o grafo. Cada aresta (u, j) considerada inconsistente terá conexão nula nas coordenadas (u, j) da matriz de adjacência, enquanto que às consistentes será atribuído o valor 1.
4. Atribuir um código único (rótulo) para cada conjunto de neurônios conectados resultante e remover grafos pequenos - com menos de 3 vértices).

Esta técnica consegue segmentar os dados a partir de informações de distância e nível de atividade. Contudo, faz uso de limiares empíricos definidos experimentalmente.

4.3.4 Agrupamento utilizando um Índice de Validação Local

Este algoritmo de agrupamento de dois níveis baseado em SOM usando um índice de validação local, proposto por (WU; CHOW, 2003), é abaixo sintetizado:

1. Treinar o SOM;
2. Aplicar o pré-processamento a seguir, antes de realizar o agrupamento sobre o SOM:
 - Primeiro, encontre os neurônios do mapa que não representam nenhum dado de entrada e elimine-os do processo de agrupamento.
 - Compute a média m_j dos vetores de dados de entrada representados por um protótipo j . Então, calcule o desvio de cada vetor de pesos w_j e m_j : $dev_j = \|w_j - m_j\|$. Calcule a média $mean_dev$ e o desvio padrão std_dev para todos os dev_j .
 - Se $dev_j > mean_dev + std_dev$, exclua o neurônio j e filtre os dados por ele representados (exclua-os do processo). Esse mecanismo elimina *outliers*.
 - Calcule as distâncias d_j entre os vetores de dados representados pelo j -ésimo neurônio e o vetor de pesos w_j . Compute a média $mean_dis_j$ e o desvio padrão std_dis_j para todas as distâncias. Calcule a média $mean_std_dis_j$ e o desvio padrão std_std_dis para todos os std_dis_j .
 - Se a distância entre os vetores de entrada x_j representados pelo j -ésimo grupo e o vetor de pesos w_j for maior que $mean_dis_j + mean_std_dis + std_std_dis$, elimine x_j . Isso pode filtrar *outliers* e ruídos.
 - Encontre o número de dados remanescentes, pertencentes ao j -ésimo grupo: num_j . Calcule a média $mean_{num}$ e o desvio padrão std_{num} para

todos os num_j .

- Se $num_j < mean_{num} - std_{num}$, exclua o neurônio j , filtrando os dados por ele representados, na tentativa de eliminar *outliers* e ruídos.

3. Agrupar o SOM usando o algoritmo hierárquico aglomerativo. O critério de junção é dado pelo índice $CDbw$ para todos os pares de grupos diretamente vizinhos². Computar o $CDbw$ para todas os vetores de dados remanescentes após filtragens, antes do processo de junção, até que somente existam dois grupos ou o processo de junção não mais possa ocorrer.
4. Encontre o particionamento ótimo dos dados de entrada pelo $CDbw$ como função que provê o número de grupos.

De acordo com os experimentos dos próprios autores, este algoritmo trabalha bem com diferentes formatos de grupos. Dado o filtro de *outliers* realizado na etapa de pré-processamento, grupos com apenas um elemento não são considerados.

4.3.5 Agrupamento considerando Topologia

Este processo de agrupamento foi proposto por (TASDEMIR; MERÉNYI, 2005). A partir de um mapa SOM treinado, o agrupamento se dá a partir de quatro fases:

- Fase 1 - Construção de um dendrograma: Em princípio, considera cada vetor de pesos como um grupo de um elemento. A proximidade mínima, Equação 2.16, é computada entre todos os grupos, e os dois mais próximos serão agrupados. Este processo é repetido até que um único grupo seja formado. Somente são incluídos nesse processo de agrupamento os protóti-

²A definição do índice $CDbw$ consta no Anexo A.

pos cujo tamanho do campo receptivo³, RF_i , satisfaz a restrição dada pela Equação 4.5.

$$|RF_i| > \mu_{RF} - \ell \cdot dev_{RF}, \quad (4.5)$$

onde $\ell > 0$ é um parâmetro de controle escalar, μ_{RF} é a média de $|RF_i|$ para todas as unidades do mapa e dev_{RF} é o desvio padrão.

- Fase 2 - Determinação de junções potenciais: Um nó no dendrograma pode ou não representar um grupo isolado. Portanto, é necessário podar o dendrograma para encontrar junções potenciais. Utilizam o valor de máximo vizinho mais próximo e de *single link*, como métricas intra e inter-grupos, respectivamente. Um nó C será rotulado para potencial junção se as distâncias do máximo vizinho mais próximo dentro de seus filhos A , B , e a distância *single link* entre eles satisfizerem o seguinte critério:

$$\frac{\max(d_{i_nnmax}(A), d_{i_nnmax}(B))}{d_{g_min}(A, B)} > \frac{level + 1}{N - 1}, \quad (4.6)$$

onde N é o número de pesos usados na construção do dendrograma e $level$ a altura da árvore.

- Fase 3 - Particionamento: As junções potenciais são então examinadas a partir da Equação 4.7 para decidir se continuam como um grupo ou serão particionadas.

$$\sum_{i \in A} |RF_i| + \sum_{j \in B} |RF_j| > \sum_{k \in (C - (A \cup B))} |RF_k| \quad (4.7)$$

³Em (HAYKIN, 1999), o termo “campo receptivo” de um neurônio, no contexto de um sistema visual, se refere à área restrita sobre a superfície retinal, que influencia as descargas daquele neurônio causadas pela luz. Entende-se, nessa metodologia, campo receptivo como sendo o peso de um neurônio.

Suponha A , B , C como potenciais grupos em uma parte da árvore. O pai C contém 2 subgrupos, A e B , que juntos totalizam 8 folhas, e 4 folhas soltas, ligadas diretamente ao pai. Se o tamanho dos campos receptivos em A e em B combinados, for menor do que o das folhas soltas, então C é mantido como um grupo. Caso contrário, haverá um particionamento. Essa decisão, segundo os autores, é tomada com o intuito de representar a densidade dos dados, melhorando as chances de encontrar grupos raros.

- Fase 4 - Junção final dos grupos: Os grupos criados nas fases anteriores passam por uma validação para verificar se precisam ser reagrupados, pois de acordo com o parâmetro *level* da fase dois, muitos grupos podem ser gerados. Dois casos são tratados: (1) junção de dois grupos com vários elementos; (2) junção envolvendo um ou dois grupos com apenas um elemento. Para apoiar essa decisão, a topologia dos dados é utilizada na criação de uma matriz de conectividade forte, $CONN$, de tamanho $M \times M$, onde M é o número de unidades do mapa, onde a conectividade forte, $CONN(i, j)$, de dois neurônios i e j é o número de exemplos de dados para os quais i ou j é o BMU e o outro o segundo BMU .

Este processo permite a descoberta de grupos com formatos arbitrários. Contudo, segundo os próprios autores, a fase de junção final dos grupos (fase 4) pode causar junções indesejadas, precisando ser otimizada.

4.4 Metodologia Proposta

A idéia central é fazer uso da propriedade de preservação de vizinhança topológica dos dados no mapeamento SOM como critério de agrupamento, utilizando-a também como forma de validação a esse processo.

Figura 4.8: Diagrama de blocos para a metodologia proposta.

A Figura 4.8 traz o diagrama de blocos que graficamente representa essa metodologia, que divide-se em 4 passos, como segue:

1. **Treinar o mapa SOM.** Nesta etapa é gerado um mapa SOM e também são calculadas as medidas Erro de Quantização (E_q), Equação 3.6, Produto Topológico (P_t), Equação 3.7, e Erro Topológico (E_t), 3.8, tentando averiguar, por meio dos valores gerados pelas mesmas, a qualidade de mapeamento aos vetores de dados de entrada.
2. **Construir um dendrograma sobre os protótipos do SOM treinado.** Esta etapa corresponde ao cálculo da proximidade mínima entre os neurônios do mapa (Equação 2.16). O objetivo aqui é encontrar os neurônios vizinhos que se agrupam pela máxima similaridade. Cada neurônio *BMU* é uma folha no dendrograma, e os demais níveis são definidos pelas ligações (junções) entre os neurônios, formando um agrupamento inicial.
3. **Avaliar potenciais grupos.** Em cada ramo do dendrograma avalia-se os potenciais grupos, no intuito de encontrar a estrutura de agrupamento subjacente aos dados, caso exista uma, determinando o número de grupos existentes na base de dados em análise. Essa etapa será detalhada à frente.

4. **Visualizar a estrutura de grupos no mapa.** A partir da estrutura de agrupamento encontrada, colore-se um mapa SOM de saída para visualização, onde cada cor representa um grupo.

Na etapa 3, para cada nível do dendrograma é, portanto, avaliado o número de grupos existentes. Uma matriz simétrica $r \times r$ de ligações (ML) é construída, onde r corresponde ao número de neurônios $BMUs$ presentes no potencial grupo (ramo do dendrograma) em avaliação. Observe a Figura 4.9 e considere, a título de ilustração, que o ramo do dendrograma em análise seja o da esquerda, que está circulado. Essa mesma figura ilustra como será a construção de ML . A diagonal principal dessa matriz é preenchida com zeros e a relação *linha* $\times$ *coluna* estabelece uma junção entre os neurônios, a qual é obtida pela Equação 4.8.

Figura 4.9: Ilustração da construção de uma matriz de ligações.

Para cada BMU , encontra-se a sua máxima distância em relação aos vetores de dados por ele representados (como se fosse estabelecido um “raio” de vizinhança a partir dessa máxima distância na representação), e a mínima distância em relação aos vetores de dados que ele não representa, tal que:

$$\min_{x_j \notin BMU_u} (d(BMU_u, x_j)) < \max_{x_i \in BMU_u} (d(BMU_u, x_i)) \quad (4.8)$$

onde $x_j \notin BMU_u$ significa que o dado x_j não é representado pelo neurônio u e $x_i \in BMU_u$ significa que o dado x_i está representado pelo neurônio u .

Em sendo a Equação 4.8 verdadeira, encontra-se o neurônio l tal que $x_j \in BMU_l$, e estabelece-se a junção entre u e l , índices de $BMUs$, em ML . Neste caso, a relação *linha* $\times$ *coluna* é considerada consistente, indicando junção, ou seja, a referida posição na matriz de ligações assume o valor 1.

A relação de ligação entre dois neurônios, no entanto, não é simétrica, de forma que pode haver divergência sobre a existência ou não de junção. Neste caso, a informação do segundo neurônio vencedor pela representação do vetor de dados é utilizada: “Se x_i tem como segundo BMU o primeiro BMU de x_j , o estabelecimento da junção dos neurônios é considerada pertinente, uma vez que os dados estão próximos no espaço de entrada dos dados”.

Neste ponto, assume-se que dados similares compartilham a mesma vizinhança, ou seja, seus respectivos primeiros e segundos $BMUs$ coincidem.

Em suma, o “agrupamento ideal” é considerado, quando, dentro cada nível do dendrograma para cada BMU , na construção da ML , salvo a diagonal principal, todas as suas posições estão preenchidas com o valor 1, ou, pela regra do segundo neurônio vencedor, vê-se a não possibilidade de junção.

A Equação 4.8 foi definida em (DING; HE, 2004), embora considerando vetores de dados de grupos diferentes, e não vetores de pesos (neurônios) e vetores de dados por ele representados, como nesta proposta. Essa função torna-se uma medida de qualidade não-supervisionada para averiguar a consistência entre os objetos vizinhos em um grupo, assumindo que para qualquer objeto em um grupo, seus vizinhos mais próximos deverão também estar no mesmo grupo, de forma que nenhum ponto x_j fora do BMU seja mais próximo dele, que qualquer outro ponto x_i por ele representado.

Esse ponto de vista também é utilizado em (JARVIS; PATRICK, 1973), que advogam que, se dois pontos, x_i e x_j forem suficientemente semelhantes, então há alta probabilidade de que eles sejam de um mesmo grupo. Logo, uma medida de similaridade pode ser aplicada, bem como um agrupamento pode ser criado.

4.5 Considerações Finais

As principais características do uso de SOM em análise de dados são sua capacidade de quantização vetorial e projeção vetorial (KASKI; KOHONEN, 1996). A quantização vetorial é feita com a projeção de N vetores de entrada para M protótipos, que representam todo o conjunto de dados original, mantendo suas características, como a topologia dos dados, mas reduzindo sua dimensão.

Com os dados dispostos no mapa em duas dimensões, esses se tornam mais simples de serem visualizados, e a organização dos vetores protótipos facilita a tarefa de agrupamento de dados, o que têm motivado a aplicação e aprimoramento deste algoritmo em aplicações de mineração de dados (VESANTO, 1997), (VESANTO; ALHONIEMI, 2000) e (JIN et al., 2004).

A partir de metodologias como as tratadas neste capítulo, o mapa SOM pode ser usado, isoladamente para agrupamento ou visualização, simultaneamente para agrupamento e visualização, ou ainda, para agrupamento por meio de visualização. Este capítulo abordou os principais trabalhos relacionados à utilização de SOM para a tarefa de análise de agrupamento, evidenciando que SOM é uma ferramenta interessante para análise exploratória de dados, em consonância com técnicas de agrupamento de dados.

5 Avaliação Experimental

“A mente que se abre a uma nova
idéia jamais voltará ao seu
tamanho original.” (*Albert
Einstein*)

O objetivo deste capítulo é apresentar, de forma experimental, a metodologia de agrupamento de dados utilizando SOM delineada no capítulo anterior, proposta central desta pesquisa. Os experimentos realizados estão divididos em três classes:

1. Um experimento com uma base de dados sintética, explicando, de forma detalhada, a execução da metodologia;
2. Experimentos com bases de dados rotuladas, para demonstrar a efetividade da proposta e compará-la à metodologia de Vesanto & Alhoniemi; e,
3. Experimentos com bases de dados reais, cujo objetivo é realmente a análise exploratória.

Os experimentos foram gerados, a partir dos recursos do ambiente computacional MatLab[®] e da ferramenta SOM Toolbox[©]¹ (VESANTO; HIMBERG J. ALHONIEMI; PARHANKANGAS, 1999).

¹O download da SOM Toolbox pode ser efetuado a partir do sítio <http://www.cis.hut.fi/projects/somtoolbox/>

A seguinte parametrização básica foi adotada em todos os experimentos realizados: mapa bidimensional, inicialização linear, treinamento em lote, formato hexagonal e função de vizinhança gaussiana. Para o número de épocas de treinamento n_e adotou-se a proposta de (VESANTO et al., 2000), de $n_e = 10 \frac{q}{n}$ épocas para a fase inicial e $n_e = 40 \frac{q}{n}$ para a fase de convergência, onde q é o número de neurônios do mapa e n o tamanho do conjunto de dados de entrada. Como taxas de aprendizado, baseado em (KASKI, 1997), adotou-se $\alpha(t) = 0,5$ na primeira fase e $\alpha(t) = 0,05$ na fase de refinamento/convergência.

5.1 Experimento Introdutório ilustrando a Proposta

Uma base de dados sintética, gerada artificialmente e de forma randômica com 270 itens de dados - que ilustra três grupos bem separados, foi escolhida para demonstrar a execução do processo de agrupamento proposto. A distribuição dos dados em classes é dada na Tabela 5.1.

Frequência	Classes Existentes
90	Classe A
90	Classe B
90	Classe C

Tabela 5.1: Distribuição da base de dados artificial em classes.

A Figura 5.1(a) ilustra a distribuição dos dados no espaço bidimensional. Um mapa SOM 6×6 foi gerado. Ao treinar um mapa SOM, este procura aproximar os neurônios ao espaço dos vetores de dados de entrada, de forma que neurônios vencedores representem esses dados.

(a)

(b)

Figura 5.1: Base de dados artificiais. Em (a), a configuração dos dados de entrada no espaço bidimensional é apresentada. Em (b), é dada a sobreposição dos neurônios da rede SOM aos dados, rotulados os dados pelo número de seu neurônio representante no mapa.

A Figura 5.1(b) apresenta essa abstração, onde os neurônios vencedores são posicionados sobre o espaço dos dados de entrada, mostrando a adaptação da rede a tais dados. Os números representam as posições dos neurônios vencedores no mapa (30 do total de 36 neurônios do mapa), e neurônios não vencedores (6 do total de 36), ou seja, que não ganharam a competição pela representação de dados, estão marcados - sem numeração, mas não serão considerados na etapa de agrupamento.

Figura 5.2: Dendrograma inicial dos 30 neurônios vencedores.

Com base no mapa SOM, um dendrograma é construído, pelo cálculo da máxima similaridade (*Single Link*) entre os neurônios vencedores, (Figura 5.2), onde no eixo x estão os *BMUs*, e no eixo y , as distâncias calculadas. Em uma abordagem *top-down*, percorre-se os níveis do dendrograma, para determinar o número de grupos existentes. Para cada ramo de um nível é calculada uma matriz de ligação, que averigua a pertinência da junção ou a necessidade de

particionamento (descer outro nível no dendrograma). A pertinência da junção é dada pela Equação 4.8 ou pela regra do segundo neurônio vencedor, como uma espécie de validação à formação final dos grupos.

(a)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
3	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
4	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	1	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0
6	0	0	1	1	1	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0
7	1	1	1	0	0	0	0	1	0	0	1	1	1	0	1	0	1	0	1	0
8	1	1	1	0	0	0	1	0	0	1	1	1	1	1	1	0	1	0	1	0
9	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0
11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0
13	1	1	1	0	0	0	1	1	0	1	1	0	0	1	0	0	0	0	0	0
14	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0
15	1	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0

(b)

Figura 5.3: Avaliação do primeiro nível do dendrograma. Em (a), é evidenciado o ramo esquerdo. Em (b), segue a matriz de ligações resultante.

As Figuras 5.3 e 5.4 mostram as matrizes de ligação dos ramos do dendrograma no primeiro nível do mesmo. Dado que a condição de parada não foi

(a)

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	1	0	0	1	0	0	1
2	1	0	0	1	0	0	1	0	0	1
3	1	1	0	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	0	1	1
5	0	0	1	0	0	1	0	1	1	0
6	1	1	1	1	1	0	1	0	1	1
7	1	1	1	1	1	1	0	1	1	1
8	1	1	1	1	1	1	1	0	1	1
9	1	1	1	1	1	1	1	1	0	1
10	1	0	1	1	1	1	1	1	1	0

(b)

Figura 5.4: Avaliação do primeiro nível do dendrograma. Em (a), é evidenciado o ramo direito. Em (b), a matriz de ligações resultante é dada.

satisfeita, uma nova iteração é iniciada, juntamente com a avaliação do segundo nível do dendrograma (Figura 5.5).

De acordo com a avaliação dos valores das matrizes de ligações dos três ramos analisados, exibidas na Figura 5.6, este é o dendrograma final, sendo realizada a poda no mesmo, considerando neste ponto, a existência de três grupos.

Figura 5.5: Dendrograma final dos neurônios vencedores.

Para bases de dados que possuem rótulos (classes definidas), como nesta base de dados, é possível rotular o mapa SOM de saída. A Figura 5.7 apresenta o mapeamento final gerado pela RNA SOM, denotando qual classe é representada por cada neurônio, bem como o número dos dados por ele representados.

A partir da análise de agrupamentos realizada pela metodologia em discussão, um mapa colorido, ou de similaridade por cor, foi gerado, e é mostrado na Figura 5.8. Cada cor evidencia um grupo encontrado, a partir dos neurônios do mapa, e dos dados por ele representados.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
7	1	1	1	0	0	0	0	1	0	0	1	1	1	0	1	1
8	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
11	1	1	1	0	0	0	1	1	0	0	0	1	1	0	1	1
12	1	1	1	0	0	0	1	1	0	0	1	0	1	0	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
15	1	1	0	0	0	0	1	0	0	0	1	1	0	0	0	1

(a)

(b)

		1	2	3	4	5	6	7	8	9	10
1	0	1	0	1	0	0	1	0	0	1	1
2	1	0	0	1	0	0	1	0	0	1	1
3	1	1	0	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	0	1	1	1
5	0	0	1	0	0	1	0	1	1	0	1
6	1	1	1	1	1	0	1	0	1	1	1
7	1	1	1	1	1	1	0	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1
9	1	1	1	1	1	1	1	1	0	1	1
10	1	0	1	1	1	1	1	1	1	0	1

(c)

Figura 5.6: Matrizes de ligação para o segundo nível do dendrograma. Em (a), a matriz condizente ao ramo de cor azul. Em (b), a matriz de ligações do ramo de cor vermelha. Em (c), a matriz do ramo de cor verde.

Figura 5.7: Mapa SOM final rotulado.

Os neurônios que não venceram na etapa de treinamento estão presentes na fase de visualização do mapa, evidenciando distância entre os grupos, ou seja, denotam que são bastante heterogêneos, bem separados entre si.

O processo de agrupamento realizado foi bastante fiel à distribuição dos dados e identificou de forma correta os grupos existentes na base de dados.

A metodologia de Vesanto & Alhoniemi foi também testada e a estrutura de agrupamento encontrada é apresentada na Figura 5.9. Pode-se perceber que para esse conjunto de dados o número de grupos encontrado é o mesmo. No entanto, considerando o conhecimento a priori sobre os dados (seus rótulos), e observando a Figura 5.8, observa-se que a estrutura encontrada/sugerida pela metodologia de análise de agrupamentos a partir de SOM proposta nesta tese é melhor definida, mais fiel à representação das classes no mapa.

Figura 5.8: Agrupamento encontrado pela metodologia proposta.

Figura 5.9: Agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.

5.2 Experimentos Comparativos sobre Bases de Dados Rotuladas

As bases de dados utilizadas nesta seção, todas de acesso público e disponíveis no repositório da Universidade da Califórnia em Irvine, “*UCI Machine Learning Repository*”, (NEWMAN et al., 1998), foram: *Iris Plants Database*, *Dermatology Database* e *Wisconsin Breast Cancer Database*. Essas bases de dados têm classes definidas e são referência na avaliação de classificadores.

De posse do número de classes, pôde-se comprovar que a metodologia identifica corretamente o número de grupos existentes e contrastar seus resultados com os produzidos pela metodologia de Vesanto & Alhoniemi.

5.2.1 Iris Plants

A base de dados *Iris*, bastante conhecida, foi criada por R. A. Fisher em 1936. Contém dados de três tipos de uma flor conhecida por Iris: Setosa, Versicolour e Virginica. Além do atributo de classificação, possui outros quatro atributos que trazem as informações de tamanho e espessura de sépala e tamanho e espessura de pétala, com todas as dimensões dadas em centímetros.

A base de dados Iris é composta por 150 amostras, sem nenhum valor ausente para seus atributos. A distribuição igualitária dos dados em classes é dada pela Tabela 5.2.

Frequência	Classes Existentes
50	Iris Setosa
50	Iris Versicolour
50	Iris Virginica

Tabela 5.2: Distribuição da base de dados *Iris* em classes.

Sabe-se que a classe Setosa é linearmente separável das classes Versicolour e Virginica, porém estas não o são entre si.

As Figuras 5.10 e 5.11 trazem as análises de agrupamentos realizadas, com tamanho de mapa diferenciado (4×4 e 5×5 , respectivamente), contudo adequados à representação dos dados, conforme avaliação da qualidade dos mapas na Tabela 5.3.

Mapa	Medidas
4×4	$E_q = 0.553$
	$E_t = 0.060$
	$P_t = 0.071$
5×5	$E_q = 0.457$
	$E_t = 0.067$
	$P_t = 0.047$

Tabela 5.3: Medidas de avaliação para os mapas da base de dados *Iris*.

Figura 5.10: Agrupamentos encontrados para a base de dados *Iris* com mapa 4×4 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.

Figura 5.11: Agrupamentos encontrados para a base de dados *Iris* com mapa 5×5 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.

Pode-se observar nas Figuras 5.10(a) e 5.10(b) que foram encontrados três grupos, muito embora, pela sobreposição dos dados, haja um grande grupo com flores do tipo Iris Virgínica e Iris Versicolor misturadas.

A Figura 5.11(a), que representa um mapa SOM de maior dimensão, indica um resultado semelhante, ou seja, três grupos, sendo dois não linearmente separáveis. Mesmo com o aumento do tamanho do mapa, percebe-se que a metodologia proposta manteve o número de grupos já encontrado, o que não ocorreu na metodologia com a qual está-se contrastando os resultados (Figura 5.11(b)).

5.2.2 Wisconsin Breast Cancer

A base de dados **Wisconsin Breast Cancer**, criada pelo Dr. W. H. Wolberg em 1991, no Hospital Universitário de Wisconsin, consiste em informações médicas ligadas ao câncer de mama. Os nove atributos biológicos existentes são, em inglês, *Clump Thickness*, *Uniformity of Cell Size*, *Uniformity of Cell Shape*, *Marginal*

Adhesion, Single Epithelial Cell, Bare Nuclei, Bland Chromatin, Normal Nucleoli e *Mitoses*, além de um atributo que define a classe do objeto.

A base de dados é composta por 699 amostras. Todos os atributos, à exceção da classe, são numéricos, e existem 16 valores ausentes nos atributos dessa base de dados, sendo que 14 destes estão presentes em registros da classe “Benigno” e 2 em registros da classe “Maligno”. Uma vez que imputação de dados foge ao escopo dessa tese, os registros com valores ausentes foram removidos da base de dados para a realização do experimento².

Nesta base de dados um tumor está classificado como sendo “Benigno” ou “Maligno”, a partir de informações colhidas a respeito das células, indicando pacientes sem a doença ou pacientes com neoplasia mamária. Após a remoção dos registros com valores ausentes, a distribuição dos dados nas classes segue como na Tabela 5.4.

Freqüência	Classes Existentes
444	Benigno
239	Maligno

Tabela 5.4: Distribuição da base de dados *Wisconsin Breast Cancer* em classes.

²Imputação de dados é a nomenclatura utilizada para denotar procedimentos que substituem os valores ausentes (desconhecidos) de um conjunto de dados por valores plausíveis. Esse assunto é coberto em (LITTLE; RUBIN, 2002) e (BATISTA, 2003).

Figura 5.12: Agrupamentos encontrados para a base de dados *Wisconsin Breast Cancer* com mapa 7×7 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi. B - Benigno; M - Maligno.

Figura 5.13: Agrupamentos encontrados para a base de dados *Wisconsin Breast Cancer* com mapa 9×9 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi. B - Benigno; M - Maligno.

A Tabela 5.5 traz os valores obtidos pelas medidas de qualidade dos mapas gerados:

Mapa	Medidas
7×7	$E_q = 0.919$
	$E_t = 0.025$
	$P_t = 0.061$
9×9	$E_q = 0.721$
	$E_t = 0.037$
	$P_t = 0.051$

Tabela 5.5: Medidas de avaliação para os mapas da base de dados *Wisconsin Breast Cancer*.

Observando as Figuras 5.12 e 5.13 averigua-se que ambas as metodologias obtiveram resultados equivalentes, e que mesmo com variação no tamanho do mapa identificaram corretamente o número de grupos existentes.

5.2.3 Dermatology Database

A base de dados *Dermatology* foi criada por I. Nilsel e colaboradores, em 1998. É formada por 33 atributos numéricos (discretos e multivariados) e 01 atributo nominal, onde estão armazenadas informações clínicas e histopatologias de pacientes a respeito de seis tipos de doenças dermatológicas: Psoríase, Seborréia, *Lichen Planus*, Pitiríase Rosa, Dermatite Crônica, e Pitiríase Vermelha.

Nesta base de dados constam 366 amostras, que possuem 8 valores ausentes no atributo idade. Esses registros com valores ausentes foram removidos da base de dados para a realização do experimento, ficando a distribuição dos como indicada na Tabela 5.6.

A Tabela 5.7, apresenta os valores produzidos pelas medidas de avaliação do mapa gerado:

Código	Frequência	Classes Existentes
1	111	Psoríase
2	60	Seborréia
3	71	<i>Lichen Planus</i>
4	48	Pitiríase Rosa
5	48	Dermatite Crônica
6	20	Pitiríase Vermelha

Tabela 5.6: Distribuição da base de dados *Dermatology* em classes.

Mapa	Medidas
4×4	$E_q = 0.827$
	$E_t = 0.008$
	$P_t = 0.062$

Tabela 5.7: Medidas de avaliação para os mapas da base de dados *Dermatology*.

Figura 5.14: Agrupamentos encontrados para a base de dados *Dermatology* com mapa 9×9 . Em (a), agrupamento dado pela metodologia proposta. Em (b), agrupamento encontrado pela metodologia de Vesanto & Alhoniemi.

Ao analisar as Figuras 5.14(a) e 5.14(b), observa-se que na primeira foram gerados seis grupos, enquanto que na segunda foram encontrados cinco grupos. Pode-se ver também, que ambos os particionamentos propostos diferem dos ró-

tulos de classes existentes. Se considerar a informação do número de grupos existente, a metodologia deste trabalho obteve melhor resultado. Isso pode ser justificado pelo fato de que para cada atributo pode ser atribuído um valor entre 0 e 3, onde 0 indica a ausência dessa característica, 3 indica alta possibilidade e 1 e 2 representam valores intermediários. Essa rotulação é bastante subjetiva, dependente do especialista, e pode influenciar na existência de uma grande proximidade, ou mesmo sobreposição, entre os grupos.

5.3 Experimentos de Análise Exploratória - Aplicações em Novas Bases de Dados Reais

Neste terceiro tipo de experimento, foram utilizadas duas bases de dados de Arqueologia, obtidas junto ao Grupo de pesquisa sobre Técnicas Analíticas Nucleares Aplicadas a Estudos Ambientais e Arqueológicos, do Instituto de Pesquisas Energéticas e Nucleares, o IPEN.

Este grupo, multidisciplinar, dedica-se à Arqueometria, um ramo da ciência que estuda as propriedades físico-químicas dos diversos vestígios de origem arqueológica. A Arqueometria faz uso de diversas técnicas físicas e químicas para obter o máximo de informação em relação aos materiais em estudo. Em especial, as técnicas de caracterização química dos vestígios cerâmicos têm sido intensivamente utilizadas, em virtude das diversas hipóteses que podem ser esclarecidas (TOYOTA et al., 2007).

O trabalho desse grupo vem contribuindo ao desenvolvimento de métodos radioanalíticos utilizando traçadores radioativos para a análise de elementos em concentrações muito pequenas (*ppm* (partes por milhão) ou *ppb* (partes por bilhão)), que são os chamados elementos traços e ultra-traços, presentes em diferentes tipos de amostras, com ênfase na análise de amostras arqueológicas, especificamente em

análise de cerâmicas e argilas de diferentes regiões como Baixo São Francisco - Sergipe, Ilha de Marajó - Pará, Manaus - Amazônia Central e dos grandes lagos do Pantanal Sul-Mato-Grossense - Mato Grosso do Sul (FELICÍSSIMA et al., 2004), (MUNITA et al., 2005) e (SANTOS, 2007). Com os resultados da composição química é possível obter informações de crucial importância à reconstituição de parte da pré-história das populações que habitaram essas regiões.

O estudo de elementos químicos que estão ao nível de traços nos fragmentos cerâmicos pode apresentar uma composição química única, podendo servir como indicadores para a identificação da fonte de matéria-prima usada nas peças. Vários métodos estatísticos multivariados vêm por eles sendo estudados para melhor caracterização desses dados, como análise de agrupamentos, componentes principais e análise discriminante, a exemplo de (GLASCOCK; NEFF; VAUGHN, 2004) e (SANTOS, 2007). A análise de agrupamentos por meio da RNA SOM emerge como mais uma ferramenta de análise exploratória para esse grupo, como evidenciam os dois experimentos abaixo descritos.

5.3.1 Base de dados de Argila do Xingu

Esse conjunto de dados versa sobre composição química e mineral de cerâmica e argila extraídas de dois depósitos explorados pela população indígena Asurini do Xingu no Estado do Pará (MUNITA et al., 2005). Foram extraídas argilas de dois depósitos dessa região em um total de 33 amostras preparadas manualmente, e coletadas em diferentes profundidades, sendo 15 para o primeiro depósito, denominado “Depósito A” e 18 para o outro, nominado “Depósito B”. A concentração dos seguintes elementos químicos foi medida: *Na*, *Lu*, *Yb*, *La*, *Th*, *Cr*, *Sc*, *Fe*, *Eu*, *Ce* e *Hf*. Dessa forma, os dados para SOM são vetores de 11 dimensões.

O objetivo dos experimentos realizados com mapas SOM foi comprovar experimentalmente a diferença entre as argilas - a existência de dois grupos distin-

tos, bem como classificar amostras de cerâmicas encontradas na região quanto à procedência, confirmando resultados experimentais obtidos por análise de discriminantes, que deram evidente diferenciação entre os depósitos (MUNITA et al., 2005), bem como o conhecimento do especialista de domínio.

Um mapa SOM 4×4 foi treinado. Os neurônios do mapa foram rotulados pelo nome do depósito (*A* ou *B*) cujos dados eles representam (Figura 5.15(a)). A linha em azul tem o objetivo de evidenciar a separação existente entre os dados. Em seguida, os dados sobre os fragmentos de cerâmica foram apresentados à rede. A Figura 5.15(b) mostra a classificação das cerâmicas, pelo posicionamento dessas amostras muito próximos ou nos mesmos neurônios que representam os dados do “Depósito *A*”.

Figura 5.15: Mapas rotulados para os dados do Xingu. Em (a), a visualização dos dados dos depósitos de argila *A* e *B* no mapa. Em (b), a classificação dos dados de cerâmica junto aos depósitos.

A análise exploratória aqui realizada - rotulação da representação dos dados por cada neurônio, em conjunto com a análise da concentração química e de análise discriminante formam uma perspectiva consistente e objetiva quanto a procedência de fragmentos de argila encontrados, contribuindo para que o arque-

ologista chegue a um consenso sobre a composição da cerâmica e a proveniência da argila de seu registro arqueológico.

5.3.2 Base de dados de Cerâmica Marajoara

Dentre as diversas grandes culturas amazônicas, uma das mais conhecidas é a Marajoara. A cultura Marajoara ocupou a Ilha do Marajó na foz do rio Amazonas entre os séculos V e XIV (SCHAAN, 1999). Este período é justamente o da Fase Marajoara, construtora dos tesos (sítios que representavam aldeias indígenas, onde hoje se encontram material arqueológico) e das urnas funerárias que lhe deram fama internacional, pela tecnologia da modelagem ceramista utilizada.

A cerâmica marajoara é tida como uma das mais belas e sofisticadas quando comparada às cerâmicas das três Américas, considerando-se suas formas bem delineadas, seus motivos variados, sua rica simbologia e sua versatilidade, pouco comum (MAGALHÃES, 2007).

Os artefatos escavados por amadores, profissionais e saqueadores, durante mais de um século, encontram-se hoje nas mãos de particulares e em diversos museus do mundo inteiro. Frente a esse fato, à compreensão da cultura dessa sociedade e à identificação de falsificação de peças da cerâmica Marajoara, é bastante importante o seu estudo.

O Museu de Arqueologia e Etnologia (MAE), da Universidade de São Paulo, possui um acervo Marajoara. A maior parte do material provém das pesquisas feitas por H. Schultz, P. P. Hilbert e Felisberto Camargo nos sítios Furinho, Cuieiras, Pacoval dos Mello, Pacoval e Fazenda São Marcos. Estas peças originalmente estavam no Museu Paulista, mas com a fusão deste com o MAE em 1989 o acervo de cerâmica Marajoara, integrado à chamada Coleção Amazônia, foi para o MAE. São 181 peças, na grande maioria cerâmicas. A coleção é constituída de

peças incompletas, à exceção de uma tanga feminina.

O museu guarda também o acervo do ex-Instituto Cultural Banco Santos, composta das coleções advindas de Graciette Lobato, proveniente da Fazenda São Marcos, e de Luiz Otávio Boulhosa, proveniente da fazenda Maravilha, ambas ao longo do rio Camutins.

A base de dados em questão contém registros de 147 fragmentos cerâmicos Marajoara que compõem o acervo do MAE-USP, cujos atributos expressam concentração de Na , K , La , Yb , Lu , U , Sc , Cr , Fe , Cs , Ce , Eu , Tb , Hf e Th determinados por meio da técnica analítica de análise por ativação com nêutrons (*Instrumental Neutron Activation Analysis*(INAA)).

Existem poucas informações sobre essas peças, uma delas é que foram encontradas na década 1930. Embora se saiba a região, não se sabe exatamente de quais tesos são oriundas. Neste sentido, são necessárias análises exploratórias, para gerar informações adicionais que orientem os arqueólogos em seu trabalho.

A análise realizada demonstra a existência de dois grupos (Figura 5.16). Embora essa informação seja bastante interessante, cabe ao arqueólogo definir se este fato se dá pela utilização de material de regiões diferentes, ou ainda, se as peças provieram de mesma região mas diferiam quanto ao uso, cotidiano ou cerimonial, pelo fato de haver diferença na tecnologia usada, tal como a temperatura de queima da argila.

Os resultados obtidos neste trabalho confirmam resultados descritos em (TOYOTA et al., 2007), que apontam a existência de dois grupos, com alguma sobreposição. Vale ressaltar que, inicialmente, os dados foram submetidos a uma análise pela distância *Mahalanobis* a fim de eliminar dados espúrios que poderiam interferir negativamente na análise. Eliminados esses dados, fez-se uma classificação preliminar pelo método de *Ward* e após, feita análise discriminante. Na

Figura 5.16: Agrupamentos encontrados para a base de dados Marajoara pela metodologia proposta.

Figura 5.17: Agrupamentos encontrados para a base de dados Marajoara pela metodologia de Vesanto & Alhoniemi.

Figura 5.18, pode-se observar o gráfico da função discriminante 1 *versus* a função discriminante 2, com um nível de confiança de 95%, segundo os autores. Os dados foram então, manualmente rotulados, de acordo com a pertinência a um dos dois grupos obtidos, nominados, respectivamente, de “Grupo 1” e “Grupo 2”.

Figura 5.18: Gráfico da análise discriminante.
Extraída de (TOYOTA et al., 2007).

Observando as Figuras 5.16 e 5.17, percebe-se que das análises de agrupamentos realizadas, a metodologia proposta obteve uma melhor representação dos grupos. Analisando também os rótulos dos neurônios e os vetores de dados representados por cada neurônio do mapa e, contrastando com os rótulos atribuídos por (TOYOTA et al., 2007), constatou-se a coincidência de 78,6% nos mesmos. Porém, muitos dos dados com rótulos diferentes estão na região de fronteira entre os grupos no mapa de similaridade por cor.

Uma vez que não se tem muitas informações sobre as peças analisadas, mesmo essa informação de incerteza sobre a classificação pode ser válida ao etnógrafo ou arqueologista em seu trabalho, uma vez que produziu-se novos conhecimentos que

auxiliarão na catalogação das peças da coleção.

5.4 Discussão Geral sobre os Resultados

Uma metodologia que utiliza o conceito de similaridade máxima baseada na preservação topológica de mapas auto-organizáveis foi proposta. Os resultados obtidos na avaliação são bastante encorajadores e validam, de forma experimental, a potencialidade da proposta.

A topologia dos dados de entrada e a asserção de que para qualquer objeto em um grupo, seus vizinhos mais próximos deverão também estar no mesmo grupo foram utilizadas no agrupamento de dados por meio de SOM. A regra do segundo *BMU* adotada, complementa o critério para o agrupamento, reforçando o uso da topologia.

Três tipos de experimentos foram gerados. O primeiro, elucidativo da metodologia. O segundo, a partir de 3 bases de dados rotuladas disponíveis no repositório do UCI, e o terceiro, de caráter exploratório. Para cada mapa utilizado foram mensurados o E_q , o E_t e o P_t , muito embora optou-se por apenas no segundo grupo de experimentos explicitá-los no texto.

Muito embora seja bastante importante essa avaliação da qualidade da quantização e da topologia no mapeamento, elas não necessariamente servem de base para, categoricamente, afirmar que um mapa está adequado ao processo de agrupamento, haja vista serem altamente dependentes da definição do tamanho do mapa. Dessas medidas, para a metodologia proposta que tem por base a preservação topológica, as mais significativas são E_t e P_t . Contudo, para grandes bases de dados o cálculo do P_t pode tornar-se inviável, por conta do seu alto custo computacional.

Muito embora as metodologias de agrupamentos baseadas em SOM assumam

que o treinamento tenha ocorrido com sucesso, ou seja, a rede elástica de neurônios se expandiu de forma a representar o espaço dos vetores de dados topologicamente, e concentrando os neurônios em regiões de maior densidade de objetos, pôde-se perceber que, para mapas muito pequenos, também é pequeno o número de grupos encontrado. Mapas muito grandes, para além de tornar o processo de aprendizagem muito lento, tendem a produzir mais grupos, podendo levar a uma estrutura de agrupamentos não existente nos dados.

Para (FLEXER, 1999), o número de neurônio do mapa SOM influencia sua aplicabilidade para agrupamento ou visualização de dados. Contudo, percebeu-se que a variação no tamanho, dentro de certo limite, mantém estável o número de grupos encontrados. Dessa forma, é bastante importante a adequação do tamanho do mapa aos dados. Essa adequação desse parâmetro é uma escolha empírica, por vezes baseada em alguma heurística, o que a torna um desafio, uma vez que o mesmo é de suma importância ao resultado final de um processo de agrupamento.

Os experimentos realizados em alguns casos foram bem melhores que a metodologia de Vesanto & Alhoniemi, a exemplo da base de dados de cerâmica Marajoara.

Ao observar a distribuição dos dados em grupos na Figura 5.18, e ver o formato dos mesmos, pode-se intuir a razão do desempenho ruim para a metodologia de Vesanto & Alhoniemi, uma vez que não são esféricos, formato este que se favorece das características do *k-means*. O fato da metodologia proposta ter encontrado o número de grupos correto, independente da variação de formato, pode ser visto como uma vantagem à mesma.

Pode-se afirmar que SOM é bastante utilizado para gerar uma organização dos dados. Utilizá-lo na análise exploratória pode ser uma tarefa simples se já

houver rótulos definidos para os dados (uma classificação prévia), pois apenas a projeção desses dados no mapa já se constitui uma informação útil.

Caso não exista esse conhecimento prévio sobre os dados, são necessárias metodologias como a que é objeto de estudo desse trabalho, para que se abram perspectivas para a descoberta de conhecimento oculto, e potencialmente relevante, a partir dos dados armazenados.

5.5 Considerações Finais

Neste capítulo, uma avaliação experimental da metodologia proposta para análise de agrupamentos a partir de SOM foi apresentada no intuito de mostrar sua efetividade, bem como de comparar com uma abordagem bastante conhecida e citada na literatura, que é a metodologia de Vesanto & Alhoniemi, que em duas fases utiliza SOM e *K-means* para realizar o agrupamento de dados.

Para além dos resultados de execução desses processos, o capítulo trata da importância da preservação topológica dos dados e do mapa SOM no agrupamento de dados, e traz uma discussão sobre a importância da qualidade desse mapeamento para o resultado final desse processo. Também foi contrastada a metodologia proposta com uma das metodologias de agrupamento descritas no Capítulo 4.

Se o tamanho do mapa não for adequado à representação dos dados, poderá haver distorções em sua representação e perda de topologia, o que implica em uma estrutura de agrupamentos inadequada, diferente à subjacente aos dados. Isto posto, pode-se afirmar que a escolha do tamanho do mapa de saída é um aspecto central na parametrização de uma RNA SOM.

Embora vê-se vantagens na metodologia proposta que não necessitam de técnicas mais elaboradas, como em SL-SOM - que se utiliza de uma técnica de

segmentação de imagens para o particionamento dos dados, e na comparação realizada, apenas com outras comparações experimentais é que poder-se-á fazer afirmações mais categóricas.

Independente da técnica de agrupamento que se apropria do mapa SOM, essa abordagem em duas fases é bastante interessante, haja vista os bons resultados gerados, a redução do custo computacional do algoritmo da segunda fase, dada a quantização realizada por SOM, e o fato de dados espúrios não influenciarem no processo de agrupamento.

6 Conclusões

“Quis mudar tudo, mudei tudo.
Agora pós-tudo, extudo... mudo!”
(*Augusto de Campos*)

O objetivo deste capítulo é apresentar as conclusões dessa pesquisa, as possibilidades de sua continuidade e os resultados nela obtidos.

6.1 Principais Considerações

A quantidade de dados disponível cresce cada vez mais, porém, seu valor está diretamente ligado à capacidade de extração de informação útil à tomada de decisões, pelos usuários finais.

Frente ao grande número de bases de dados existentes e carentes de análise, e frente às limitações humanas de, sem auxílio de recursos computacionais apropriados, realizar essa análise, emerge uma nova necessidade, por ferramentas analíticas que utilizem técnicas capazes de transformar grandes quantidades de dados em informações significativas, ou seja, em conhecimento capaz de ser utilizado de forma direta.

Motivado por esse cenário, este trabalho discutiu análise de agrupamentos e mapas auto-organizáveis, interligando esses conceitos na formação de uma

metodologia de geração de grupos que utiliza informações de topologia dos dados nessa tarefa. Embora haja outras metodologias análogas, esta apresenta-se como uma alternativa viável à análise exploratória de dados.

O primeiro capítulo trouxe a contextualização, a motivação e os objetivos almejados nessa pesquisa. No Capítulo 2 teve-se por objetivo fornecer um panorama geral sobre a área de análise de agrupamento de dados e seus desafios.

A análise de agrupamentos é uma importante ferramenta para exploração de dados não rotulados, por meio da construção de hierarquias ou pela formação de grupos homogêneos. É um processo iterativo, formado por vários passos específicos, como pré-processamento, desenvolvimento e aplicação de um algoritmo de agrupamento e pós-processamento (validação e avaliação).

A rede neural SOM e conceitos a ela adjacentes foram abordados no Capítulo 3, uma vez que a mesma é utilizada como primeira fase na metodologia de análise de agrupamentos aqui proposta.

Embora metodologias de agrupamento de dados utilizando SOM venham sendo aplicadas como ferramentas de visualização de dados de alta dimensionalidade, os resultados sempre estarão sujeitos a algum tipo de interpretação, e para análise de agrupamentos faz-se necessário algum tipo de processamento sobre o mapa topológico gerado.

É fato que a visualização dos grupos existentes em uma base de dados também é uma verificação crucial dos resultados de um processo de agrupamentos. Como encontrar grupos a partir de um mapa SOM e a verificação visual desses grupos foram assuntos do Capítulo 4, que apresentou a descrição conceitual da metodologia de agrupamentos proposta nesse trabalho.

Sabe-se, no entanto, que as ferramentas de visualização existentes podem não ser tão claras aos usuários finais da aplicação, principalmente para conjunto de

dados multidimensionais. Logo, é importante a aplicação de procedimentos adicionais como contagem de frequência dos dados nos grupos, análise de correlação entre atributos e validação do agrupamento, para uma análise de agrupamento mais significativa.

A análise de agrupamento divide os dados em grupos com propósito de sumariação ou de melhorar a compreensão sobre os dados. Verificou-se na pesquisa que são muitas as suas aplicações, como agrupamento de documentos para busca, encontrar genes e proteínas de funcionalidade similar, como forma de compressão de dados, entre tantas outras.

No quinto capítulo, a metodologia idealizada foi aplicada, em três diferentes tipos de experimentos, a fim de validá-la empiricamente, e de forma comparativa. Os resultados finais dos experimentos foram bastante satisfatórios, mostrando a efetividade da proposta.

Essa pesquisa constatou que SOM tem-se firmado como uma ferramenta interessante para análise exploratória de dados, em consonância com técnicas de agrupamento de dados e de visualização. Independente da técnica de agrupamento usada, essa abordagem em duas fases tem como vantagem a geração de bons resultados, a redução do custo computacional do algoritmo na segunda fase - pela quantização do SOM, bem como por dados espúrios não influenciarem no processo de agrupamento.

Conclui-se afirmando que, muito embora a análise de agrupamentos tenha uma longa história e que um grande número de técnicas de agrupamento venham sendo desenvolvidas nas comunidades de Estatística, Reconhecimento de Padrões, Mineração de Dados, entre outras áreas, os desafios significativos ainda prevalecem, como: encontrar conjuntos com tamanhos, formas e densidades diferentes, tratamento de ruídos e dados espúrios (pré-processamento), e como determinar o

número de grupos existentes.

A seleção de uma solução final de agrupamento de dados ainda requer um substancial julgamento de um especialista de domínio, e pode ser considerada subjetiva. Isto posto, as pesquisas nessa área são promissoras e quaisquer contribuições, mesmo bastante pontuais, são bastantes relevantes, pelo auxílio na geração de novos conhecimentos sobre uma base de dados.

6.2 Principais Contribuições

Esta tese focaliza-se principalmente na análise exploratória de dados, que pode ser vista como uma fase inicial em um processo de mineração de dados, que objetiva a compreensão ou geração de hipóteses sobre os dados. As principais contribuições desse trabalho seguem abaixo arroladas:

- A proposição de uma nova metodologia de agrupamento de dados a partir de SOM pautada em duas premissas centrais: (1) para qualquer objeto em um grupo, seus vizinhos mais próximos deverão também estar no mesmo grupo, e (2) dados similares compartilham a mesma vizinhança, ou seja, seus respectivos primeiros e segundos *BMUs* coincidem. Utiliza-se, portanto, de informações das topologias dos dados e do mapa SOM.
- O melhor desempenho, nos testes realizados, da proposta de análise de agrupamentos em comparação à de Vesanto & Alhoniemi, bastante difundida, que realiza agrupamentos a partir de SOM.
- A aplicação de SOM e análise de agrupamentos em dados reais, explorando bases de dados de arqueologia pode ser vista como uma contribuição de ordem prática, atestando a importância da análise exploratória, e comprovando a eficiência da metodologia proposta frente a outras técnicas analíticas

utilizadas nessas bases de dados.

- Uma vez que há uma confusão sobre a utilização de SOM em agrupamento de dados, a explanação do Capítulo 4, que discute as três formas de utilização de SOM em análise exploratória de dados, pode ser considerada, pela síntese, uma contribuição. As três formas são: (1) como algoritmo de agrupamento quando há conhecimento a priori sobre o número de grupos existentes, e assume-se esse valor como número de neurônios do mapa; (2) como um algoritmo que indica a tendência de agrupamentos, constatada visualmente; e, (3) como fase inicial em uma metodologia de agrupamento de dados, seguida da aplicação de alguma técnica para esse fim.

Finalmente, com base nos resultados experimentais obtidos, pôde-se constatar a viabilidade da proposta, que pode ser aplicada em diversas áreas do conhecimento para compreensão de seus dados.

6.3 Trabalhos Futuros

Como trabalhos futuros a esta pesquisa, pode-se citar:

- Dar continuidade aos estudos sobre critérios para agrupamentos, no intuito de encontrar soluções alternativas e passíveis de aplicações reais, na determinação do número de grupos presentes em uma base de dados, de forma a propor melhorias na metodologia desenvolvida neste trabalho.
- Verificar o desempenho da metodologia proposta com outras bases de dados, que apresentem características diferentes (formas, tamanhos e densidades) das presentes nas bases de dados utilizadas nos experimentos.
- Realizar novos experimentos, principalmente em bases de dados reais, onde haja pouco conhecimento prévio e a necessidade da produção de novos “in-

sights” para suporte à tomada de decisão, como as coletadas por sensores e biosensores. Ou ainda, estender a aplicação da metodologia advogada nessa tese para outros domínios de bases de dados, como o de imagens médicas.

- Ampliar a avaliação comparativa para outras metodologias de análise de agrupamento, de forma a contrastar o desempenho dessa proposta frente aos resultados obtidos por outras metodologias não executadas no âmbito da pesquisa dessa tese.
- A partir de um mapa segmentado, como o de similaridade por cores, definir critérios e realizar análises para compreender quais atributos contribuem para a maior proximidade entre os grupos, de forma a melhor entender a formação de um grupo e a relação desse grupo com o restante da base de dados, principalmente, com os grupos presentes em seu entorno. Ainda, investigar quais são os atributos que os grupos têm em comum, e que podem contribuir sobremaneira para a existência de sobreposições.

Uma vez que a análise exploratória de dados emprega grande variedade de técnicas visuais e quantitativas, no intuito de maximizar a obtenção de informações ocultas em sua estrutura, descobrir atributos importantes para o aprendizado, testar hipóteses assumidas ou escolher modelos matemáticos, entre outras, são atividades de interesse. Ainda, considerando a necessidade emergente de descoberta de conhecimento em bases de dados e o caráter multidisciplinar que a ciência vem assumindo, vê-se um amplo espaço para futuras pesquisas nessa área.

6.4 Publicações

Ao longo do período de doutoramento junto ao Grupo ICONE (Grupo de Inteligência Computacional, Modelagem e Neurocomputação)¹ da EP-USP, foram publicados os seguintes trabalhos, resultantes dos trabalhos de pesquisa e cooperação inerentes à produção dessa tese:

- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D. M. Using Self-Organizing Maps and NearestNeighbor to Data Clustering. In: Proceedings of the ICAI08 The 2008 International Conference on Artificial Intelligence, Las Vegas, USA, 2008.
- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D. M., SASSI, R. J. A Data Mining Methodology to Support Customer Relationship Management. In: Proceedings of the V CONTECSI - The 2008 International Conference on Information Systems and Technology Management, São Paulo/SP, 2008.
- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D. M. Clustering de Dados de Sensores a partir de Mapas Auto-organizáveis. In: Anais do IBERSENSOR 2006 5o. Congreso Iberoamericano de Sensores, Montevideo, 2006.
- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D. M. Análise de Agrupamentos utilizando Mapas Auto-organizáveis em Agricultura de Precisão. In: Anais do Congresso Brasileiro de Agricultura de Precisão, São Pedro SP. ConBAP 2006, 2006.

¹O sítio do Grupo está disponível no endereço: <http://www.lsi.usp.br/icone/>

Anexo A – Algoritmos de Validação

Este anexo descreve os algoritmos de validação de agrupamento citados no texto da tese.

A.1 Índice Davies-Bouldin

O índice *Davies-Bouldin* (DB) (DAVIES; BOULDIN, 1979), Equação A.1, indica a similaridade entre grupos. Pode ser utilizada na validação de partições de dados e na comparação relativa entre diferentes segmentações de um conjunto de dados.

$$DB = \frac{1}{C} \sum_{k=1}^C \max_{l \neq k} \left\{ \frac{S_c(Q_k) + S_c(Q_l)}{d_{ce}(Q_k, Q_l)} \right\}, \quad (\text{A.1})$$

onde C é o número de grupos, $S_c(Q_k)$ representa a distância intra-grupos do agrupamento Q_k , baseada na distância para o centróide, (A.2):

$$S_c(Q_k) = \left\{ \frac{1}{N_k} \sum_{j=1}^{N_k} |c_k - x_j|^p \right\}^{\frac{1}{p}}, \quad (\text{A.2})$$

onde $x_j \in Q_k$, N_k é o número de dados no agrupamento Q_k e $c_k = \frac{1}{N_k} \sum_{x_i \in Q_k} x_i$.

Já $d_{ce}(Q_k, Q_l)$, que representa a distância entre grupos, também baseada em centróide, é dada por:

$$d_{ce}(Q_k, Q_l) = \left\{ \sum_{k=1}^d |c_k - x_k|^p \right\}^{\frac{1}{p}}, \quad (\text{A.3})$$

onde d é a dimensão do vetor x_k .

O índice DB é independente do número de grupos e do algoritmo de particionamento, o que o torna bom candidato à avaliação desses algoritmos.

A.2 Índice CDbw

O índice $CDbw$ (*Composing Density Between and With Clusters*), proposto por (HALKIDI; VAZIRGIANNIS, 2002), baseia-se em distâncias intra e inter-grupos.

Para cada conjunto de dados dividido em c grupos é definido um conjunto de pontos representativos $V'_i = \{v_{i1}, v_{i2}, \dots, v_{iri}\}$ para o agrupamento i , onde r_i representa o número de pontos de representação para o grupo i . Calcula-se o desvio padrão $stdev(i)$ para cada componente p do agrupamento i , como segue:

$$stdev^p(i) = \sqrt{\sum_{k=1}^{n_i} \frac{(x_k^p - m_i^p)^2}{(n_i - 1)}}, \quad (\text{A.4})$$

onde n_i é o número de dados constantes no i -ésimo agrupamento e m_i a média dos dados do grupo i .

A densidade intra-grupos é definida por:

$$Intra_den(c) = \frac{1}{c} \sum_{i=1}^c \sum_{j=1}^{r_i} density(v_{ij}), \quad c > 1 \quad (\text{A.5})$$

O termo $density(v_{ij})$ é definido como $density(v_{ij}) = \sum_{l=1}^{n_i} f(x_l, v_{ij})$, tal que x_l pertence ao i -ésimo agrupamento, v_{ij} é o j -ésimo ponto representativo do i -ésimo grupo, n_i é o tamanho do i -ésimo grupo, e $f(x_l, v_{ij})$ é definida por:

$$f(x_l, v_{ij}) = \begin{cases} 1, & \text{se } \|x_l - v_{ij}\| \leq \|stdev(i)\| \\ 0, & \text{caso contrário.} \end{cases} \quad (\text{A.6})$$

A densidade entre grupos é definida como segue:

$$Inter_den(c) = \sum_{i=1}^c \sum_{j=1, j \neq i}^c \frac{\|close_rep(i) - close_rep(j)\|}{\|stdev(i)\| + \|stdev(j)\|} density(v_{ij}), c > 1 \quad (A.7)$$

onde $close_rep(i)$ e $close_rep(j)$ representam o par de pontos de representação mais próximos entre os agrupamentos i e 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 é um vetor de entrada pertencente ao i -ésimo ou j -ésimo agrupamento, n_i e n_j representam, respectivamente, o número de dados existentes nos grupos i e j , e $f(x_k, v_{ij})$ é dado por:

$$f(x_k, v_{ij}) = \begin{cases} 1, & \text{se } \|x_k - v_{ij}\| \leq (\|stdev(i) + stdev(j)\|)/2 \\ 0, & \text{caso contrário.} \end{cases} \quad (A.8)$$

A separabilidade dos grupos é calculada por meio de:

$$Sep(c) = \sum_{i=1}^c \sum_{j=1, j \neq i}^c \frac{\|close_rep(i) - close_rep(j)\|}{\|1 + Inter_den(c)\|} density(v_{ij}), c > 1 \quad (A.9)$$

O índice CDbw é dado por:

$$CDbw(c) = Intra_den(c) \times Sep(c) \quad (A.10)$$

O índice *CDbw* ressalta características geométricas de cada agrupamento, representada pelos vetores representativos de cada agrupamento, sendo eficiente, portanto, para agrupamentos de formas arbitrárias. No caso de SOM, ao invés de centróides, os vetores de pesos são utilizados.

Referências Bibliográficas

AGRESTI, A. *Categorical Data Analysis*. 2a. ed. Nova York, USA: Wiley-Interscience, 2002.

ALHONIEMI, E. *Unsupervised Pattern Recognition Methods for Exploratory Analysis of Industrial Process Data*. Tese (PhD. Thesis) — Helsinki University of Technology, Espoo, Finlândia, Dezembro 2002.

AROUS, N.; ELLOUZE, N. On the search of fitness evaluation of a self-organizing map by means of organization measures. *Proceedings of the International Symposium on Computational Intelligence and Intelligent Informatics (ISCHII'03)*, Nabeul, Tunísia, p. 75–79, Maio 2003.

BAEZA-YATES, R. A.; FRAKES, W. B. Introduction to data structures and algorithms related to information retrieval. Prentice-Hall, Inc., San Diego, CA, USA, p. 13 – 27, 1992.

BAÇÃO, F.; LOBO, V.; PAINHO, M. Self-organizing maps as substitutes for k-means clustering. *Proceedings of the International Conference on Computational Science (ICCS'2005)*, Springer-Verlag, Berlin, p. 476–483, 2005.

BARLOW, H. B. Unsupervised learning. *Neural Computation*, v. 1, p. 295–311, 1989.

BATISTA, G. E. A. P. A. *Pré-processamento de Dados em Aprendizado de Máquina Supervisionado*. Tese (Doutorado) — USP - Universidade de São Paulo, São Carlos - SP, Maio 2003.

BAUER, H. U.; HERRMANN, M.; VILLMANN, T. Neural maps and topographic vector quantization. *Neural Networks*, v. 12, n. 4–5, p. 659–676, 1999. Disponível em: <citeseer.ist.psu.edu/bauer99neural.html>.

BAUER, H. U.; PAWELZIK, K. R. Quantifying the neighborhood preservation of self-organizing feature map. *IEEE Transactions on Neural Networks*, v. 3, n. 4, p. 570–579, 1992.

BEAR, M. F.; CONNORS, B. W.; PARADISO, M. A. *Neurociências: Desvendando o Sistema Nervoso*. 2a. ed. Porto Alegre: Artmed Editora S.A., 2002.

BERKHIN, P. *Survey of Clustering Data Mining Techniques*. Relatório Técnico, Accrue Software, San Jose, CA, 2002.

- BEUCHER, S.; LANTUÉJOUL, C. Use of watersheds in contour detection. *Proceedings of the International Workshop Image Processing, Real-Time Edge and Motion Detection/Estimation*, CCETT/INSA/IRISA, Rennes, França, p. 2.1 – 2.12, Setembro 1979.
- BISWAS, G.; WEINBERG, J. B.; FISHER, D. H. Iterate: A conceptual clustering algorithm for data mining. *IEEE Transactions on Systems, Man, and Cybernetics: Applications and Reviews*, IEEE Computer Society, Washington, DC, USA, v. 28-C, p. 219 – 230, Maio 1998.
- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D.-M. Análise de agrupamentos utilizando mapas auto-organizáveis em agricultura de precisão. *Anais do Congresso Brasileiro de Agricultura de Precisão (ConBAP'2006)*, ABAP, São Pedro - SP, Junho 2006.
- BOSCARIOLI, C.; SILVA, L. A.; HERNANDEZ, E. D.-M. Clustering de dados de sensores a partir de mapas auto-organizáveis. *Anais do 5º Congresso Iberoamericano de Sensores - IBERSENSOR'2006*, CyTED, Montevideo - Uruguai, p. 1 – 5, Setembro 2006.
- BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. *Redes Neurais Artificiais - Teoria e Aplicações*. Rio de Janeiro: LTC, 2000.
- BRUN, M.; SIMA, C.; HUA, J.; LOWEY, J.; CARROLL, B.; SUH, E.; DOUGHERTY, E. R. Model-based evaluation of clustering validation measures. *Pattern Recognition*, Elsevier Science, Oxford, v. 40, n. 2, p. 807 – 824, Março 2007.
- BUSSAB, W. O.; MIAZAKI, . S.; ANDRADE, D. F. *Introdução à Análise de Agrupamentos*. São Paulo, Julho 1990.
- CHEN, K.; LIU, L. *Towards Finding Optimal Partitions of Categorical Datasets*. Relatório Técnico, Syracuse University, EUA, 2003.
- CORMACK, R. M. A review of classifications. *Journal of the Royal Statistical Society*, Series A, n. 134, p. 321–353, 1971.
- COSTA, J. A. F. *Classificação Automática e Análise de Dados por Redes Neurais Auto-Organizáveis*. Tese (Doutorado) — UNICAMP - Universidade Estadual de Campinas, Campinas - SP, Dezembro 1999.
- COSTA, J. A. F.; NETTO, M. L. A. Segmentação do som baseada em particionamento de grafos. *Anais do VI Congresso Brasileiro de Redes Neurais*, São Paulo, p. 451–456, Junho 2003.
- COTTRELL, M.; BODT, E. A kohonen map representation to avoid misleading interpretations. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN'1996)*, D-Facto Public, Bruges, Bélgica, p. 103 – 110, Abril 1996.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, Abril 1979.

DING, C. H. Q.; HE, X. K-nearest-neighbor consistency in data clustering: Incorporating local information into global optimization. *Proceedings of the ACM Symposium on Applied Computing - SAC'04*, ACM, Nicosia, Cyprus, p. 584 – 589, Março 2004. Disponível em: <<http://doi.acm.org/10.1145/967900.968021>>.

EL-HAMDOUCHI, A.; WILLETT, P. Hierarchic document classification using ward's clustering method. *Proceedings of the 9th Annual International Conference on Research and Development in Information Retrieval - SIGIR'86*, ACM SIGIR, Palazzo dei Congressi, Pisa, Itália, p. 149–156, Setembro 1986. Disponível em: <<http://doi.acm.org/10.1145/253168.253200>>.

ERWIN, E.; OBERMAYER, K.; SCHULTEN, K. Self-organizing maps: Ordering, convergence properties and energy functions. *Biological Cybernetics*, Springer-Verlag, v. 67, p. 47–55, 1992.

ERWIN, E.; OBERMAYER, K.; SCHULTEN, K. Self-organizing maps: Stationary states, metastability and convergence rate. *Biological Cybernetics*, Springer-Verlag, v. 67, p. 35–45, 1992. Disponível em: <<http://citeseer.ist.psu.edu/350419.html>>.

EVERITT, B. S.; LANDAU, S.; MORVEN, L. *Cluster Analysis*. 4a. ed. Londres: Hodder Arnold Publishers, 2001.

EVERITT, B. S.; RABE-HESKETH, S. *The Analysis of Proximity Data*. Londres: Hodder Arnold Publishers, 1997.

FAUSSET, L. *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Nova Jersey, USA: Prentice-Hall, 1994.

FELICÍSSIMA, M. P.; PIREAUX, J. J.; DEMORTIER, G.; RODRIGUES, U. P. F. Estudos arqueométricos de cerâmicas indígenas pré-coloniais das lagoas castelo e vermelha, localizadas no pantanal sul-mato-grossense. *Canindé - Revista do Museu de Arqueologia de Xingó*, MAX/UFS, Sergipe, n. 4, p. 325 – 368, 2004.

FLEXER, A. On the use of self-organizing maps for clustering and visualization. *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery - PKDD'99*, Springer-Verlag, Londres, UK, p. 80 – 88, Junho 1999.

GIBSON, D.; KLEINBERG, J.; RAGHAVAN, P. Clustering categorical data: An approach based on dynamical systems. *The Very Large Database Journal*, Springer Verlag, v. 8, n. 2, p. 222 – 236, Março 2000.

GLASCOCK, M. D.; NEFF, H.; VAUGHN, K. J. Instrumental neutron activation analysis and multivariate statistics for pottery provenance. *Hyperfine Interactions*, Kluwer Academic Publishers, Netherlands, v. 154, n. 1 – 4, p. 95 – 105, Novembro 2004.

GOODHILL, G. J.; SEJNOWSKI, T. J. Quantifying neighbourhood preservation in topographic mappings. *Proceedings of the 3rd Joint Symposium on Neural Computing*, University of California, San Diego and California Institute of Technology, Pasadena, CA, p. 61–82, Junho 1996.

GRINSTEIN, G.; TRUTSCHL, M.; CVEK, U. High-dimensional visualizations. *Proceedings of the 7th Data Mining Conference - KDD*, São Francisco, Califórnia, 2001.

GRIRA, N.; CRUCIANU, M.; BOUJEMAA, N. Unsupervised and semi-supervised clustering - a brief survey. *A Review of Machine Learning Techniques for Processing Multimedia Content*, p. 1–12, outubro 2004. Disponível em: <<http://www-rocq.inria.fr/~crucianu/src/BriefSurveyClustering.pdf>>.

HAIR, J. F.; TALHAM, R. L.; ANDERSON, R. E.; BLACK, W. C. *Multivariate Data Analysis*. 5a. ed. Nova Jersey, USA: Prentice Hall, 1998.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. *Journal of Intelligent Information Systems*, Kluwer Academic Publishers, Hingham, MA, USA, v. 17, n. 2-3, p. 107–145, Julho 2001. ISSN 0925-9902.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part i. *SIGMOD Record*, v. 31, n. 2, p. 40–45, 2002.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity checking methods: Part ii. *SIGMOD Record*, v. 31, n. 3, p. 19–27, 2002.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multi representatives. *Proceedings of Hellenic Conference on Artificial Intelligence - SETN'02*, Thessaloniki, Grécia, Abril 2002.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 1a. ed. Nova York: Morgan Kaufmann Publishers, 2001. ISBN 1-55860-489-8.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. 2a. ed. Porto Alegre: Bookman, 1999.

HEBB, D. O. *The Organization of Behavior*. Nova York: John Wiley and Sons, Inc., 1949.

HIMBERG, J. *From Insights to Innovations: Data Mining, Visualization, and User Interfaces*. Tese — Helsinki University of Technology, Espoo, Finlândia, Novembro 2004.

HÖPPNER, F.; KLAWONN, F.; KRUSE, R.; RUNKLER, T. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*. Nova York, USA: Wiley Publishers, 1999.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Nova Jersey, USA: Prentice Hall, 1988. (PH Advanced Reference Series). ISBN 0-13-022278-X.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. *ACM Computing Surveys*, v. 31, n. 3, p. 264–323, Setembro 1999.

JARVIS, R. A.; PATRICK, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, IEEE Computer Society, Washington, DC, USA, C-22, n. 11, p. 1025 – 1034, Novembro 1973. ISSN 0018-9340. Disponível em: <<http://dx.doi.org/10.1109/T-C.1973.223640>>.

JIN, H.; SHUM, W. H.; LEUNG, K. S.; WONG, M. L. Expanding self-organizing map for data visualization and cluster analysis. *Information Sciences*, Elsevier, Inc., v. 163, p. 157–173, Outubro 2004.

KASKI, S. *Data Exploration using Self-organizing Map*. Tese (PhD. Thesis) — Helsinki University of Technology, Espoo, Finlândia, Março 1997.

KASKI, S.; KANGAS, J.; KOHONEN, T. Bibliography of self-organizing maps (som) papers: 1981 - 1997. *Neural Computing Surveys*, v. 1, p. 102–350, Maio 1998. Disponível em: <citeseer.ist.psu.edu/104693.html>.

KASKI, S.; KOHONEN, T. Winner-take-all networks for physiological models of competitive learning. *Neural Networks*, v. 7, p. 973–984, 1994.

KASKI, S.; KOHONEN, T. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. *Proceedings of the 3rd International Conference on Neural Networks in the Capital Markets*, World Scientific, Singapura, p. 498–507, 1996.

KOHONEN, T. Analysis of a simple self-organizing process. *Biological Cybernetics*, Springer-Verlag, v. 44, n. 2, p. 135–140, 1982.

KOHONEN, T. *Self-Organization and Associative Memory*. 3a. ed. Nova York, USA: Springer-Verlag, 1989.

KOHONEN, T. *Self-Organizing Maps*. 3a. ed. Nova York, USA: Springer-Verlag, 2001.

KOTSIANTIS, S. B.; PINTELAS, P. E. Recent advances in clustering: A brief survey. *Transactions on Information Science and Applications*, WSEAS, v. 1, n. 1, p. 73–81, 2004. ISSN 1790-0832.

KUMAR, V.; TAN, P.-N.; STEINBACH, M. *Introduction to Data Mining*. 1a. ed. Boston: Addison-Wesley, 2005.

- LAMPINEN, J.; OJA, E. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, v. 2, n. 2–3, p. 261–272, 1992.
- LATTIN, J.; CARROLL, J. D.; GREEN, P. E. *Analyzing Multivariate Data*. Cole: Thomson Books, 2003. (Duxbury Applied Series).
- LITTLE, R. J.; RUBIN, D. B. *Statistical Analysis with Missing Data*. 2a. ed. Nova York, USA: John Wiley and Sons, 2002.
- MAGALHÃES, M. P. *A Cultura Marajoara*. Novembro 2007. Disponível em: <<http://recantodasletras.uol.com.br/artigos/737106>>.
- MALSBURG, C. von der. Network self-organization. Academic Press Professional, Inc., San Diego, CA, USA, p. 421–432, 1990.
- MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Pfychometrika*, Springer, Nova York, v. 50, n. 2, p. 159 – 179, Junho 1985.
- MUNITA, C. J. A. S.; SILVA, M. A.; SILVA, F. A.; OLIVEIRA, P. M. S. Archaeometric study of clay deposits from the indigenous land of the asurini do xingu. *Instrumentation Science & Technology*, Taylor & Francis, Londres, UK, v. 33, n. 2, p. 161 – 173, Março 2005. Disponível em: <<http://dx.doi.org/10.1081/CI-200048058>>.
- NEWMAN, D. J.; HETTICH, S.; BLAKE, C. L.; MERZ, C. J. *UCI - Repository of Machine Learning Databases*. 1998. Disponível em: <<http://archive.ics.uci.edu/ml/>>.
- OJA, M.; KASKI, S.; KOHONEN, T. Bibliography of self-organizing map (som) papers: 1998-2001 addendum. *Neural Computing Surveys*, v. 3, n. 1–156, 2002. Disponível em: <citeseer.ist.psu.edu/683884.html>.
- OLIVEIRA, M. C. F.; LEVKOWITZ, H. From visual data exploration to visual data mining: A survey. *IEEE Transactions On Visualization And Computer Graphics*, v. 9, n. 3, p. 378–394, Julho-Setembro 2003.
- PAL, S. K.; SRIMANI, P. K. Neurocomputing: Motivations, models, and hybridization. *IEEE Computer*, v. 29, n. 3, p. 24–28, Março 1996.
- PALMER, C. R.; FALOUTSOS, C. Electricity based external similarity of categorical attributes. *Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Seul, Coréia, p. 120 – 126, Abril 2003.
- PARENTE, V.; ZANINI, A.; HERNANDEZ, E. D.-M.; SOUZA, R. C. Utilização de redes neurais como ferramenta quantitativa na regulação tarifária do setor elétrico. *Anais do III Congresso Brasileiro de Regulação dos Serviços Públicos Concedidos*, Associação Brasileira das Agências de Regulação (ABAR), Gramado - RS, Maio 2003.

PERES, S. M. *Dimensão Topológica e Mapas Auto Organizáveis de Kohonen*. Tese (Doutorado) — UNICAMP - Universidade Estadual de Campinas, Campinas - SP, Setembro 2006.

PÖLLÄ, M.; HONKELA, T.; KOHONEN, T. Bibliography of self-organizing maps (som) papers: 2002-2005 addendum. *Neural Computing Surveys*, 2007. Forthcoming.

PÖLZLBAUER, G. Survey and comparison of quality measures for self-organizing maps. *Proceedings of the 5th Workshop on Data Analysis (WDA'04)*, Elfa Academic Press, Slovakia, p. 67–82, Junho 2004.

QIU, B.-Z.; ZHANG, X.-Z.; SHEN, J.-Y. Grid-based clustering algorithm for multi-density. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, p. 18–21, agosto 2005.

RITTER, H.; SCHULTEN, K. On the stationary state of kohonen's self-organizing sensory mapping. *Biological Cybernetics*, Springer, Berlin/Heidelberg, v. 54, n. 2, p. 99–106, Junho 1986.

RITTER, H.; SCHULTEN, K. Convergence properties of kohonen's topology conserving maps: Fluctuations, stability, and dimension selection. *Biological Cybernetics*, Springer-Verlag, v. 60, p. 59–71, 1988.

ROCHA, H. V.; LACHI, R. L. *Aspectos Básicos de Clustering: Conceitos e Técnicas*. Relatório Técnico IC-05-003, UNICAMP - Universidade Estadual de Campinas, Campinas, fevereiro 2005.

RUMELHART, D. E.; ZIPSER, D. Feature discovery by competitive learning. *Cognitive Science*, v. 9, p. 75–112, 1985. Disponível em: <<http://cognitn.psych.indiana.edu/rgoldsto/cogsci/Rumelhart.pdf>>.

SANTOS, J. O. *Estudos Arqueométricos de sítios arqueológicos do baixo São Francisco*. Tese (Doutorado) — Instituto de Pesquisas Energéticas e Nucleares, São Paulo - SP, 2007.

SASSI, R. J. *Uma arquitetura híbrida para descoberta de conhecimento em bases de dados: teoria dos rough sets e redes neurais artificiais mapas auto-organizáveis*. Tese (Doutorado) — USP - Universidade de São Paulo, São Paulo - SP, Novembro 2006.

SCHAAN, D. P. Cultura marajoara: História e iconografia. *Resgate da Cultura Material e Iconográfica do Pará*, Sebrae/MPEG, Belém-PA, I - Arte Rupestre e Cerâmica, 1999.

SCHEAFFER, R. *Categorical Data Analysis*. Relatório Técnico. NCSSM Statistics Leadership Institute, Carolina do Norte, EUA, 1999.

TASDEMIR, K.; MERÉNYI, E. Considering topology in the clustering of self-organizing maps. *Proceedings of the 5th Workshop On Self-Organizing Maps (WSOM'2005)*, Paris, França, p. 439–446, Setembro 2005.

THEODORIDIS, S.; KOUTROUMBAS, K. *Pattern Recognition*. 3a. ed. Londres: Academic Press, 2005.

TOYOTA, R. G.; MUNITA, C. J. A. S.; NEVES, E. G.; DEMARTINI, C. C. Preliminary study of the effect of temper in marajoara ceramics. *Anais do 1o Simpósio Latino Americano sobre Métodos Físicos e Químicos em Arqueologia, Arte e Conservação de Patrimônio Cultural.*, MASP - Museu de Arte de São Paulo “Assis Chateaubriand”, São Paulo, p. 1 – 6, Junho 2007.

ULTSCH, A. Self-organizing neural networks for visualization and classification. *Information and Classification*, Springer-Verlag, Dortmund, Alemanha, p. 307–313, 1992.

ULTSCH, A. Knowledge extraction from self-organizing neural networks. *Information and Classification*, Springer, Berlin, p. 301–306, 1993.

ULTSCH, A.; SIEMON, H. P. Kohonen’s self-organizing feature maps for exploratory data analysis. *Proceedings of the International Neural Network Conference - INNC'90*, Kluwer Academic Publishers, Dordrecht, Netherlands, p. 305 – 308, Julho 1990.

URIARTE, E. A.; MARTÍN, F. D. Topology preservation in som. *Transactions on ENgineering, Computing and Technology*, v. 15, p. 187–190, 2006.

VAPNIK, V.; BEN-HUR, A.; HORN, D.; SIEGELMANN, H. Support vector clustering. *Journal of Machine Learning Research*, v. 2, p. 125–137, 2001. Disponível em: <citeseer.ist.psu.edu/hur01support.html>.

VESANTO, J. *Data Mining Techniques Based on the Self-Organizing Map*. Dissertação (Msc. Thesis) — Helsinki University of Technology, Espoo, Finlândia, Maio 1997.

VESANTO, J. Som-based data visualization methods. *Intelligent Data Analysis*, Elsevier Science, v. 3, n. 2, p. 111–126, 1999.

VESANTO, J. *Using SOM in Data Mining*. Dissertação (Licenciate Thesis) — Helsinki University of Technology, Espoo, Finlândia, Abril 2000.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, v. 11, n. 3, p. 586–600, Maio 2000.

VESANTO, J.; ALHONIEMI, E.; HIMBERG, J.; PARHANKANGAS, J. *SOM Toolbox for Matlab 5*. Relatório Técnico A57, Helsinki University of Technology, Espoo, Finlândia, Abril 2000.

VESANTO, J.; HIMBERG J. ALHONIEMI, E.; PARHANKANGAS, J. Self-organizing map in matlab: The som toolbox. *Proceedings of the Matlab DSP Conference 1999*, Digital Signal and Image Processing - DSP Group, Espoo, Finlândia, p. 35–40, Novembro 1999. [Http://citeseer.ist.psu.edu/vesanto00selforganizing.html](http://citeseer.ist.psu.edu/vesanto00selforganizing.html).

VILLMANN, T.; DER, R.; HERRMANN, M.; MARTINETZ, T. M. Topology preservation in self-organizing feature maps: Exact definition and measurement. *IEEE Transactions on Neural Networks*, IEEE Computer Society, Washington, DC, USA, v. 8, n. 2, p. 256 – 266, Março 1997.

WU, S.; CHOW, T. W. S. Self-organizing-map based clustering using a local clustering validity index. *Neural Processing Letters*, Kluwer Academic Publishers, Netherlands, v. 17, p. 253–271, 2003.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645–678, Maio 2005.

ZHANG, Y.; FU, A. W.; CAI, C. H.; HENG, P. Clustering categorical data. *Proceedings of the 16th International Conference on Data Engineering*, San Diego, USA, p. 305 – 316, Fevereiro 2000.

ZUCHINI, M. H. *Aplicações de Mapas Auto-organizáveis em Mineração de Dados e Recuperação de Informação*. Dissertação (Mestrado) — UNICAMP - Universidade Estadual de Campinas, Campinas - SP, Setembro 2003.