AUTÔMATOS CELULARES GENERALIZADOS COMO MODELOS DE INFLUÊNCIA PARA AGRUPAMENTOS DE DADOS E INTERAÇÕES SOCIAIS

Eduardo José Aguilar Alonso

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. Luis Alfredo Vidal de Carvalho, D.Sc.

Prof. Adilson Elias Xavier, D.Sc.

Prof. Carlos Eduardo Pedreira, Ph.D.

Prof. Lúcia Maria de Assumpção Drummond, D.Sc.

Prof. Roseli Suzi Wedemann, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

JUNHO DE 2008

AGUILAR ALONSO, EDUARDO JOSÉ

Autômatos Celulares Generalizados como Modelos de Influência para Agrupamentos de Dados e Interações Sociais. [Rio de Janeiro] 2008

XI, 90 p. 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia de Sistemas e Computação, 2008)

Tese – Universidade Federal do Rio de Janeiro, COPPE

1. Atômato celular
2. Modelos de influência.
3. Agrupamentos de dados.
4. Interações sociais.

I. COPPE/UFRJ II. Título (série)

Aos três anjos que a vida me deu: Teresita, Vanessa e Ayumi.

Agradecimentos

Ao longo de cinco anos de trabalho muitas pessoas contribuíram diretamente ou indiretamente, de maneira consciente ou não, na consecução do meu objetivo. A cada uma delas, de acordo com o meu estilo, manifestei minha gratidão assim que eu compreendi o valor da sua contribuição. Entretanto, sinto a necessidade de fazer explícito o meu reconhecimento àqueles que mais diretamente e ativamente participaram na elaboração deste trabalho.

O ponto de partida encontra-se no estímulo e apoio dos meus dois orientadores: Luis Alfredo Vidal de Carvalho e Valmir Carneiro Barbosa. Para mim é uma honra ter trabalhado junto a dois excelentes pesquisadores e ter contado com o conselho de duas personalidades evoluídas e refinadas.

Toda aventura em terras inexploradas implica em riscos. Felizmente, durante o caminho todo eu tive o privilégio da companhia dos meus três grandes amigos e colegas: Nicolau Maluf, Rogerio Salvini e Bruno Osiek. A atenção e a paciência que me brindaram nos meus intermináveis devaneios me ajudaram a distinguir as musas das sereias.

O meu trabalho foi enriquecido pelas sugestões e conselhos de Raúl Donángelo e de Lúcia Drummond.

A criatividade e a inspiração me visitaram provavelmente atraídas pela paz e o aconchego que minha esposa Elize Ayumi Hayashi fornece permanentemente ao meu lar. A sua ajuda e apoio incondicional foram indispensáveis para que eu pudesse concluir esta etapa da minha vida.

Eu quero agradecer ao CNPq pelo apoio financeiro, que permitiu a realização desta tese.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

AUTÔMATOS CELULARES GENERALIZADOS COMO MODELOS DE INFLUÊNCIA PARA AGRUPAMENTOS DE DADOS E INTERAÇÕES SOCIAIS

Eduardo José Aguilar Alonso

Junho / 2008

Orientadores: Valmir Carneiro Barbosa
Luís Alfredo Vidal de Carvalho

Programa: Engenharia de Sistemas e Computação

Os sistemas complexos estão compostos por muitos elementos de caraterísticas semelhantes com interações mútuas tipicamente não-lineares. Em alguns desses sistemas, a origem da não-linearidade pode ser modelada considerando que cada elemento possui uma propriedade, que chamaremos influência, cujo valor é co-determinante da intensidade das interações de um elemento com os restantes, e que por sua vez, o histórico de interações modifica o valor da influência. Os autômatos celulares são uma das ferramentas conceituais mais empregadas na modelagem de sistemas complexos. No presente trabalho desenvolvemos dois autômatos celulares generalizados (ACG), um deles determinístico e o outro estocástico, para aplicar o conceito de influência ao problema de detecção de agrupamentos de dados e à dinâmica das relações sociais. O ACG determinístico é o embasamento de um novo algoritmo para detecção de agrupamentos em grandes bancos de dados. O conceito de vizinhança, inerente aos ACGs, é a única medida de distância empregada; por este motivo, os dados não requerem nenhum pré-processamento. A localidade do processamento da informação dos ACGs confere ao algoritmo uma aptidão intrínseca para sua paralelização. A escalabilidade do algoritmo com respeito à quantidade de dados a considerar deriva-se do fato que as próprias células que compõem o ACG representam subconjuntos. A capacidade de auto-organização dos ACG permite que os agrupamentos emerjam sem a necessidade de parâmetros. São apresentados resultados da aplicação do algoritmo a bancos de dados sintéticos e naturais. Finalmente, o ACG estocástico aplicado à dinâmica das relações sociais resulta num modelo que reproduz as tendências observadas em bancos de dados criminalísticos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

GENERALIZED CELLULAR AUTOMATA AS MODELS OF INFLUENCE IN DATA CLUSTERING AND SOCIAL INTERACTIONS

Eduardo José Aguilar Alonso

June / 2008

Advisors: Valmir Carneiro Barbosa
Luis Alfredo Vidal de Carvalho

Department: Systems Engineering and Computer Science

Complex systems are large sets of elements with similar characteristics interacting in a non-linear way with each other. In some of these systems, the source of non-linearity can be modeled considering that each element has a property, which we call influence, whose value is co-determinant of the intensity of the interactions of an element with the others, and that in turn, the history of interactions alters the value of the influence. Cellular automata are common tools employed for modeling complex systems. In this work we develop two generalized cellular automata (GCA), one of them deterministic and the other, stochastic, to apply the concept of influence to the problem of data clustering, and to the dynamics of social relations. The deterministic GCA is the core of a new clustering algorithm for large databases. The concept of neighborhood, as part of the cellular automaton definition, is the only measure of distance employed and for this reason, the data do not require any pre-processing. Since GCAs perform local processing of information, the algorithm is naturally suited to be parallelized. The scalability of the algorithm with respect to the amount of data to be processed is assured since the cells that make up the GCA themselves represent subsets of data. Self-organization, a property found in many GCAs, is responsible for the emergence of clusters without need of tuning parameters. Several tests, performed both on synthetic and natural databases serve to illustrate the performance of the algorithm. Finally, the stochastic GCA applied to the dynamics of social relations results in a model that reproduces the observed trends in databases of criminal activities.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os sistemas complexos não possuem ainda uma definição formal, sendo isto uma conseqüência da ausência de um tratamento analítico dentro da teoria geral dos sistemas dinâmicos [58]. Devido a essa falta de definição formal, os sistemas complexos são identificados por meio de algumas caraterísticas gerais. Por exemplo, para ILIACHINSKI [31] todos os sistemas complexos se compõem de um grande número de partes interconectadas com iterações mútuas tipicamente não lineares (pp. 612). Exemplos de sistemas complexos podem ser achados em áreas e escalas tão diferentes quanto uma galáxia e o cérebro humano. São inúmeros os trabalhos que explicitamente ou implicitamente definem o seu objeto de estudo como um sistema complexo. Centros como o SANTA FE INSTITUTE[1] foram criados com o propósito de estudar estes sistemas. Devido à ampla diversidade de áreas onde estudos desta natureza podem ser aplicados, as abordagens multidisciplinares são as mais freqüentes.

As definições ou delimitações de um sistema complexo, como por exemplo a citada acima, herdam deste a dificuldade de explicar ou capturar a essência embutida num conjunto de elementos para que estes exibam fenômenos que qualitativamente classificamos como complexos. Nem todo conjunto de elementos interagindo apresenta os sinais de complexidade que interessam. Apenas aqueles sistemas que apresentam padrões espaço-temporais que LANGTON [44] classifica na *fronteira do caos* são os interessantes. A principal dificuldade para fazer um estudo analítico destes sistemas está no fato de que não tem sido possível predizer a sua evolução numa escala dada de observação baseados somente nos fenômenos observáveis nessa escala. Também tem-se fracassado, na maioria dos casos, em predizer a formação de padrões globais conhecendo os detalhes da dinâmica das interações entre os elementos que compõem o sistema. Por este motivo os fenômenos observáveis na escala global de um sistema complexo são chamados *emergentes* [36]. Nestas condições as simulações computacionais aparecem como a alternativa mais eficiente para deixar em evidência os fenômenos emergentes a partir da especificação da

[1] http://www.santafe.edu

topologia de interconexões e interações entre os elementos constituintes.

Na década de 1940, VON NEUMANN considerou o problema de criar uma entidade artificial com capacidade de se auto-reproduzir. Por sugestão de ULAM, VON NEUMANN começou a trabalhar com um conjunto de entidades simples sediadas num reticulado, estruturalmente e funcionalmente idênticas. Ele chamou a cada uma dessas entidades de *célula*, e batizou o conjunto de *autômato celular* (CA). Com esta abordagem, ao problema da auto-reprodução se acrescentou a necessidade de inventar regras de comportamento e interação para as entidades individuais com o objetivo de fazer *emergir* o comportamento coletivo da auto-reprodução. A solução que VON NEUMANN apresentou aparece descrita no livro *Theory of Self-Reproducing Automata* [78]. A solução original de VON NEUMANN foi aprimorada por outros autores [75, 16]. Provavelmente a maior contribuição derivada da pesquisa de VON NEUMANN seja o autômato celular como ferramenta para estudar a relação entre simples comportamentos individuais e os complexos comportamentos coletivos emergentes. Autores como WOLFRAM [81] destacam o potencial dos autômatos celulares para modelar sistemas complexos, assim como a sua capacidade computacional. Nesse sentido TOFFOLI [76] mostra que o espaço de soluções abrangido pelos autômatos celulares é o mesmo que o das equações diferenciais, e portanto eles podem ser usados como uma alternativa para criar e analisar modelos em física. Uma análise dos aspectos formais por trás da computabilidade dos autômatos celulares pode ser achada em [19].

No próximo capítulo será apresentada uma definição e descrição detalhada de um autômato celular; entre as propriedades que compõem um CA está o conceito de *vizinhança*. Cada célula que compõe o CA possui uma vizinhança que consiste de um conjunto de células do CA que são chamadas de *vizinhas*. Uma célula interage somente com aquelas células que são suas vizinhas; portanto, para conhecer a evolução de uma célula somente é necessário dispor da informação contida na célula em questão e as suas vizinhas. Destarte, a evolução de cada célula que compõe o CA pode ser determinada em forma simultânea para todas elas. O potencial para o processamento paralelo de informação acrescenta o atrativo dos CAs para analisar modelos de sistemas complexos [8].

Alguns sistemas complexos se auto-organizam em configurações quase-estáveis, nas quais pequenas perturbações podem levar a mudanças na configuração global. Essas mudanças, que não podem ser caraterizadas a partir da intensidade da perturbação, podem ficar localizadas em pequenas regiões do sistema ou atingí-lo integramente. Exemplos deste tipo de comportamento são ubíquos na natureza: avalanches de neve, terremotos, incêndios florestais, etc. O trabalho de BAK, TANG e WEISENFELD [4] foi o primeiro a chamar este comportamento de *criticalidade auto-organizada*. Nesse trabalho foi empregado um autômato celular para mostrar numericamente a presença de uma lei de potência associada à distribuição de freqüência dos tamanhos das mudanças na configuração global. Os livros de BAK [3] e JENSEN [35] fazem uma ampla revisão dos trabalhos re-

ferentes ao estudo de sistemas complexos que evoluem para configurações criticamente auto-organizadas.

No entanto, nem todos os sistemas complexos se auto-organizam numa configuração global crítica. Por exemplo, BARABÁSI e ALBERT [7] reencontram a lei de potência na distribuição de freqüência do grau dos nós de redes tais como a *World Wide Web*. Este tipo de sistema se auto-organiza para chegar numa configuração robusta; segundo os autores, duas caraterísticas devem ser levadas em conta ao se modelar tais redes: os nós se integram gradativamente à rede, e a sua integração está regida por uma *preferência* ou *bias*. Esta preferência se corresponde reciprocamente com a *influência* que nós muito conectados têm sobre a escolha dos novos para estabelecer conexões. Ainda que o modelo original não considere somente informação local nas interações entre nós, é relevante para este trabalho a observação que alguns elementos do sistema possuem uma influência maior do que outros; em outros termos, ainda que estruturalmente os elementos que compõem o sistema sejam idênticos, as interações entre eles aparecem condicionadas por um atributo que, neste caso, tem o seu valor ajustado em decorrência da evolução do sistema. A presença deste atributo modulador das interações é determinante na formação das configurações globais auto-organizadas. Outros exemplos de redes auto-organizadas aparecem descritas no livro de BARABÁSI [6]. Note-se que redes podem ser modeladas de muitas maneiras, entre elas por autômatos celulares.

Autômatos celulares têm sido usados com sucesso na modelagem de crescimento urbano. Por exemplo, BARREDO *et al.* [9] utilizam um autômato celular estocástico para predizer o crescimento da cidade de Dublin (Irlanda). Na primeira parte do trabalho, os autores justificam o porquê do uso dos autômatos celulares; dentre as justificativas destacamos duas: ainda que tudo possa estar relacionado com tudo, coisas vizinhas estão mais relacionadas que as distantes, e a evolução de uma porção de terra depende da influência à qual esteja submetida. No primeiro ponto é ressaltada a importância de considerar vizinhanças para especificar a dinâmica do uso da terra em cidades. O segundo ponto coincide com a consideração feita por BARABÁSI na importância de reconhecer que há assimetrias nas interações entre elementos. WARD *et al*. [80] apresentam um trabalho sobre crescimento urbano onde o cenário é Queensland, Austrália. Eles também empregam um autômato celular estocástico para fazer a modelagem e retomam o conceito de influência.

Uma sociedade composta por indivíduos é outro exemplo de sistema complexo que pode se beneficiar de simulações computacionais [55]. Por exemplo, DUFFY e OCHS [22] usaram um modelo computacional para pesquisar as condições para a validade de um modelo teórico proposto por Kiyotaki e Wright; segundo esse modelo, o dinheiro como meio de intercâmbio é um fenômeno emergente, que deve sua aparição à auto-organização das relações comerciais entre indivíduos. Estudos desta natureza fazem parte de uma jovem área chamada *econofísica*. NOWAK *et al*. [52] empregam um autômato celular para estu-

dar a inter-relação entre decisões econômicas e fenômenos sociais como a influência na formação de opiniões. Num estudo prévio, NOWAK *et al*. [56] empregam mais uma vez um autômato celular para estudar a emergência e a estabilidade da *identidade de grupo*. A influência que um indivíduo recebe de seus pares é chamada de pressão. Variações nesta pressão afetam a coesão do grupo assim como a sua estabilidade perante informações vindas de fora do grupo. Nessa mesma linha de raciocínio, a personalidade de um indivíduo – composta de sentimentos e pensamentos – emerge como resultado de dois mecanismos básicos: a sincronização dos estados internos de indivíduos em interação e a auto-organização de pensamentos e sentimentos com respeito à auto-imagem do indivíduo [57].

A influência do grupo sobre um indivíduo – seja este de qualquer espécie – está na base da proposta de JUANICO, MONTEROLA e SALOMA [37] para quem a *alelomimese*[2] – a tendência a comportar-se como os outros – é um mecanismo geral para a emergência de agrupamentos. Este mecanismo é reconhecido em algumas espécies de insetos sociais [13]. Ele é usado XU, CHEN e HE [84] para construir um autômato celular estocástico que sedia formigas virtuais, as quais imitam o comportamento das formigas reais de agrupar-se para ganhar proteção.

No presente trabalho desenvolvemos dois modelos gerais de autômato celular, um deles determinístico e o outro estocástico, com o objetivo explorar o conceito de influência que permeia os trabalhos supracitados. O autômato celular determinístico é usado num novo algoritmo de detecção de agrupamentos de dados que não usa nenhuma medida de distância além da implícita na definição da vizinhança de uma célula. Este algoritmo aproveita a auto-organização das células para fazer emergir os agrupamentos dos dados. O segundo autômato celular é projetado para incluir, além da influência assimétrica entre indivíduos, os mecanismos gerais dos seres vivos de adaptação e acomodação. Em ambos os casos são apresentados resultados computacionais que respaldam a pertinência de tais modelos.

No próximo capítulo é feito um resumo do embasamento teórico necessário para a compreensão dos modelos propostos. No terceiro capítulo é apresentado o autômato celular determinístico e o algoritmo para a detecção de agrupamentos de dados; são incluídos os resultados computacionais sobre diferentes bancos de dados. No quarto capítulo é introduzido o autômato celular estocástico usado para modelar a dinâmica de grupos de indivíduos – de qualquer espécie – levando em conta o mecanismo de influência; também é apresentado um resultado computacional que põe em evidência a tendência à auto-organização e que se corresponde com os dados de bancos de dados reais sobre comportamentos delituosos. Finalmente, no quinto capítulo são resumidas as conclusões e os trabalhos futuros.

[2]No texto original é usado o termo allelomimesis. Usamos uma adaptação ao português.

Capítulo 2

Fundamentos Teóricos

O presente capítulo tem por objetivo introduzir as definições daqueles conceitos empregados no desenvolvimento do trabalho.

2.1 Autômato celular

A partir do trabalho de WOLFRAM, "*Statistical mechanics of cellular automata*" [82], os autômatos celulares começaram a ser aplicados massivamente ao estudo de sistemas complexos. No entanto, como foi indicado na Introdução, sua primeira aplicação foi o estudo da auto-reprodução feito por VON NEUMANN [78]. O livro de ILIACHINSKI [31] e o livro de WOLFRAM [81] são as duas principais obras de referência.

Numa primeira aproximação, um autômato celular (AC) é um objeto abstrato composto por duas componentes intrinsecamente ligadas. A primeira delas é um reticulado *(lattice) regular, discreto* e *infinito* que constitui o *universo* ou a *estrutura* do *espaço subjacente* do AC. A segunda é um *autômato finito* que se repete de maneira idêntica em cada nó do reticulado. Cada um desses autômatos é denominado *célula*. Uma célula dada c manterá um intercâmbio de mensagens com um número finito de outras células que, ao serem consideradas como um conjunto, constituem a *vizinhança* da célula c. A comunicação de c com a suas vizinhas é *local, determinística, uniforme e síncrona*. Ela determina a evolução global do sistema ao longo de *passos* num *tempo discreto*. Vejamos a seguir a definição formal do autômato celular clássico.

2.1.1 Definição de autômato celular

Um autômato celular d-dimensional $\mathscr{A}$ é um quarteto $(\mathbb{Z}^d, \mathscr{S}, \mathscr{N}, \delta)$, onde:

- $\mathbb{Z}^d$ representa o espaço com d dimensões de números inteiros; cada ponto desse espaço – também chamado de *reticulado* – está ocupado por uma célula,

- $\mathscr{S}$ é um conjunto finito, cujos elementos representam os estados possíveis de cada célula de $\mathscr{A}$,
- $\mathscr{N}$ é um subconjunto finito de $\mathbb{Z}^d$, chamado *vizinhança* de $\mathscr{A}$, onde cada elemento $\mathbf{n} \in \mathscr{N}$ deve ser interpretado como as coordenadas relativas das células vizinhas; ou seja, dada uma célula c localizada em $\mathbf{p} \in \mathbb{Z}^d$, a sua vizinhança $\mathscr{N}(c)$ é o conjunto de células localizadas em $\mathbf{p}+\mathbf{n}$; por simplicidade, suporemos que $(0,...,0) \in \mathscr{N}$;
- $\delta : \mathscr{S}^{k+1} \to \mathscr{S}$, é a *regra local de transição* de $\mathscr{A}$, onde k é a quantidade de células em $\mathscr{N}$.

O conjunto $\mathscr{S}$ pode possuir estados especiais s, chamados *inativos -quiescent-* para os quais se verifica que $\delta(s,\ldots,s) = s$.

2.1.2 *Configuração* ou *estado global* de $\mathscr{A}$

Dado um autômato celular $\mathscr{A} = (\mathbb{Z}^d, \mathscr{S}, \mathscr{N}, \delta)$ podemos lhe associar uma aplicação $c_t^{\mathscr{A}} : \mathbb{Z}^d \longmapsto \mathscr{S}$ que fornece para cada instante $t \geq 0$ o estado de todas e cada uma das células que compõem $\mathscr{A}$. Essa aplicação é conhecida como *configuração* ou *estado global* de $\mathscr{A}$ no instante t.

2.1.3 Vizinhanças

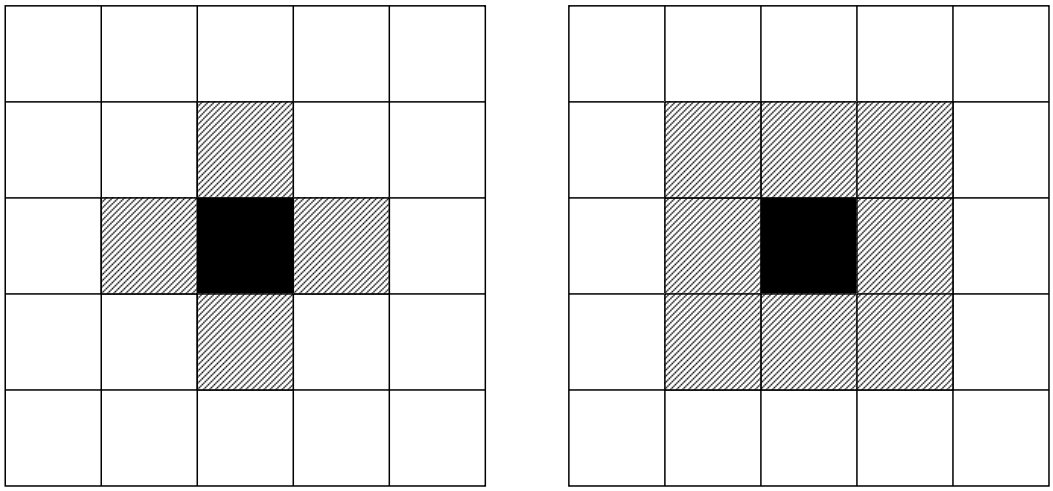
Consideremos um autômato celular $\mathscr{A} = (\mathbb{Z}^d, \mathscr{S}, \mathscr{N}, \delta)$. A vizinhança de uma célula c, $\mathscr{N}(c)$ é o conjunto de células do reticulado que determinarão localmente o próximo estado de c. Nos ACs clássicos, a vizinhança é finita e geometricamente uniforme, ainda que ela possa estar composta por qualquer conjunto finito ordenado. As duas vizinhanças mais simples são a de VON NEUMANN e a de MOORE, as quais definimos a seguir.

Seja c uma célula no ponto $\mathbf{z}$ de $\mathbb{Z}^d$, com $\mathbf{z} = (z_1,\ldots,z_d)$; consideremos as normas $\| \mathbf{z} \|_1 = \sum_{i=1}^{d} |z_i|$, e $\| \mathbf{z} \|_\infty = \max\{|z_i| \,|\, i \in \{1,\ldots,d\}\}$. As definições das duas vizinhanças supra-citadas são as seguintes:

- vizinhança de VON NEUMANN: $\mathscr{N}_{VN}(c) = \{\mathbf{x} | \mathbf{x} \in \mathbb{Z}^d, \| \mathbf{z}-\mathbf{x} \|_1 \leq 1\}$, com uma ordem dada,
- vizinhança de MOORE: $\mathscr{N}_M(c) = \{\mathbf{x} | \mathbf{x} \in \mathbb{Z}^d, \| \mathbf{z}-\mathbf{x} \|_\infty \leq 1\}$, com uma ordem dada.

Na Figura 2.1 estão representadas ambas as vizinhanças no caso particular de um reticulado plano sobre $\mathbb{Z}^2$.

Ambas vizinhanças estão compostas pelos vizinhos mais próximos; no entanto elas podem ser facilmente generalizadas para incluir células mais distantes. Para isso é suficiente retomar as definições anteriores e introduzir o parâmetro $r \in \mathbb{N}^*$, chamado *raio da*



(a) Von Neumann (b) Moore

Figura 2.1: Vizinhanças clássicas num reticulado plano.

vizinhança. Destarte, as definições das vizinhanças de VON NEUMANN e MOORE com raio r assumem as seguintes expressões:

- vizinhança de VON NEUMANN com raio r: $\mathscr{N}_{VN}^{r}(c) = \{\mathbf{x} \,|\, \mathbf{x} \in \mathbb{Z}^d, \| \mathbf{z} - \mathbf{x} \|_1 \leq r\}$, com uma ordem dada,
- vizinhança de MOORE com raio r: $\mathscr{N}_{M}^{r}(c) = \{\mathbf{x} \,|\, \mathbf{x} \in \mathbb{Z}^d, \| \mathbf{z} - \mathbf{x} \|_\infty \leq r\}$, com uma ordem dada.

2.1.4 Regras locais de transição

A regra local de transição determina qual será o próximo estado de uma célula c a partir dos estados das células que compõem a vizinhança. Há uma forma comum de especificar essa regra por meio de uma tabela. Consideramos um AC unidimensional tal que a vizinhança (com $r = 1$) de c inclui a célula à esquerda, *lc*, e a célula à direita, *rc*. A Tabela 2.1 apresenta um exemplo de regra local. WOLFRAM [82] introduziu uma nomenclatura para representar sistematicamente as regras locais usando para isso uma interpretação lexicográfica do número binário gerado a partir da coluna de *bits* com o próximo estado de c. Assim, na Tabela 2.2 está a explicação de como gerar a representação de WOLFRAM para o exemplo de regra local citada acima, à qual possui o identificador 54. Decorre disto que os AC unidimensionais, binários, com vizinhança de cardinalidade 3 terão ao todo $2^8 = 256$ regras diferentes. Todos os ACs obtidos considerando cada uma dessas regras locais, foram estudados exaustivamente chegando-se à conclusão que nenhum deles tem um comportamento dinâmico global "interessante": ao longo de muitos passos todos os ACs convergem para a configuração global na qual todas as células têm o mesmo estado (ver [8]).

Tabela 2.1: Exemplo de regra local, onde $lc[t]$ e $rc[t]$ representam respectivamente os estados no passo t das vizinhas à esquerda e à direita da célula c; a última coluna representa o próximo estado da célula c.

$lc[t]$	$c[t]$	$rc[t]$	$c[t+1]$
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	0

Tabela 2.2: Exemplo de representação da regra 54 segundo WOLFRAM. Na primeira linha, cada grupo de três *bits* representa o grupo de estados das vizinhas (nos extremos) e da célula considerada (*bit* central). A segunda linha linha contem o próximo estado da célula considerada; estes bits são usados na terceira linha para gerar o número identificador da regra.

000	001	010	011	100	101	110	111
0	1	1	0	1	1	0	0
0×2^0	1×2^1	1×2^2	0×2^3	1×2^4	1×2^5	0×2^6	0×2^7

2.1.5 Reticulados

Diversos tipos de reticulados têm sido considerados dependendo do objetivo do estudo. Assim, abordagens teóricas dos AC consideram geralmente reticulados infinitos com dimensões $d \geq 1$. Do ponto de vista teórico tem-se percebido a existência de um salto entre o tratamento de AC unidimensionais e os de dimensões maiores. A maior parte dos resultados demonstrados para $d = 1$, são indecidíveis para $d \geq 2$ [19].

Em aplicações onde ACs são usados como modelos, têm sido empregados outros tipos de reticulados. Em simulações computacionais somente podem ser considerados reticulados finitos; no entanto, o tamanho pode ser escolhido de maneira que os efeitos de borda não tenham efeito sobre as simulações. Há um compromisso entre o tamanho do reticulado e o número de passos necessários para que os efeitos das bordas se apresentem numa região reduzida no centro do espaço.

Uma alternativa para evitar efeito de borda é a de considerar um reticulado onde, para cada dimensão, as bordas estão fundidas. Assim, no caso unidimensional, o reticulado forma um anel; para $d = 2$, se todas as dimensões fundem as suas bordas gera-se um reticulado toroidal.

Alguns trabalhos supõem que as bordas estão formadas por células que não mudam o seu estado, sendo este geralmente o estado inativo (ou *quiescent*) [37]. Este tipo de

reticulado possui a chamada *condição de borda nula* (*null boundary condition*).

2.2 Generalizações do autômato celular

O autômato celular definido na seção anterior, devido à sua simplicidade, foi extensamente estudado, tanto analiticamente como usando simulações. Como resultado desses trabalhos, propriedades tais como reversibilidade e capacidade computacional ganharam teoremas, mas em geral eles só se aplicam para o caso unidimensional.

As generalizações surgem como conseqüência de mudar a escolha das componentes do quarteto que compõe $\mathscr{A}$. SARKAR [66] faz uma revisão sistemática das diversas variantes que têm sido estudadas. A lista a seguir não é exaustiva; ela serve apenas para dar um contexto aos modelos desenvolvidos nos dois capítulos seguintes.

2.2.1 Variantes de $\mathscr{S}$

Nas abordagens teóricas dos ACs supõe-se que o conjunto de estados possíveis $\mathscr{S}$ para uma célula é finito, e geralmente inclui um estado especial chamado *inativo (quiescent)*. Na definição clássica do AC, todas as células possuem o mesmo $\mathscr{S}$. No entanto, têm sido estudados ACs nos quais $\mathscr{S}$ varia de uma célula para outra; eles são chamado de ACs *poligênicos*.

Nas aproximações teóricas $\mathscr{S}$ é finito, freqüentemente $\mathbb{Z}_m$ (conjunto dos inteiros módulo m); outros estudos consideram que o estado de uma célula pode assumir um inteiro arbitrário, sendo portanto $\mathscr{S}$ infinito. Em aplicações de modelagem de um fenômeno físico ou biológico, o estado de uma célula pode assumir valores contínuos. Este tipo de generalização, ainda que bastante comum, vai contra a idéia de que no AC tudo - espaço, tempo, estados - tenha valores discretos. Segundo FEYNMAN [23], à importância de considerar valores discretos se relaciona uma consideração mais profunda: a modelagem baseada em variáveis contínuas implica na concentração local de uma quantidade infinita de informação. A abordagem clássica empregando equações diferenciais se baseia em variáveis contínuas. Entretanto, TOFFOLI [76] mostra que o espaço de soluções que os ACs abrangem é igual àquele das equações diferenciais. Parece portanto, que modelagens baseadas em ACs, considerando variáveis e tempo discretos, descrevem com a mesma exatidão o comportamento de um sistema dinâmico, ainda que a informação processada localmente seja finita.

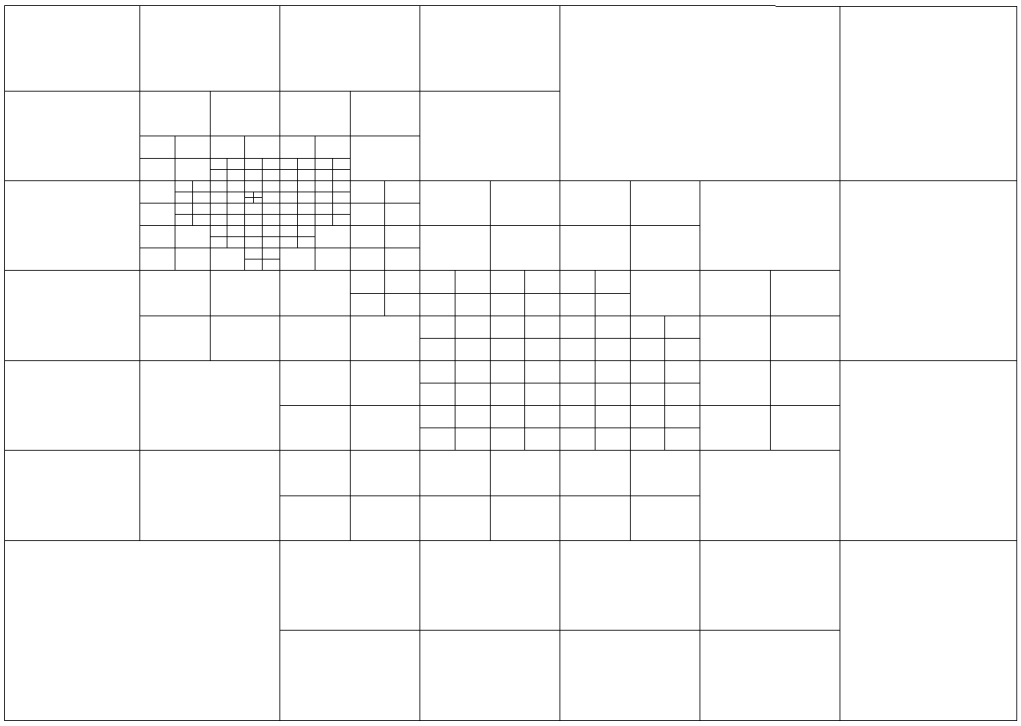
2.2.2 Variantes do reticulado

O AC apresentado inicialmente por VON NEUMANN possuía um reticulado bidimensional infinito. Em estudos teóricos a suposição de reticulados infinitos é a mais comum.

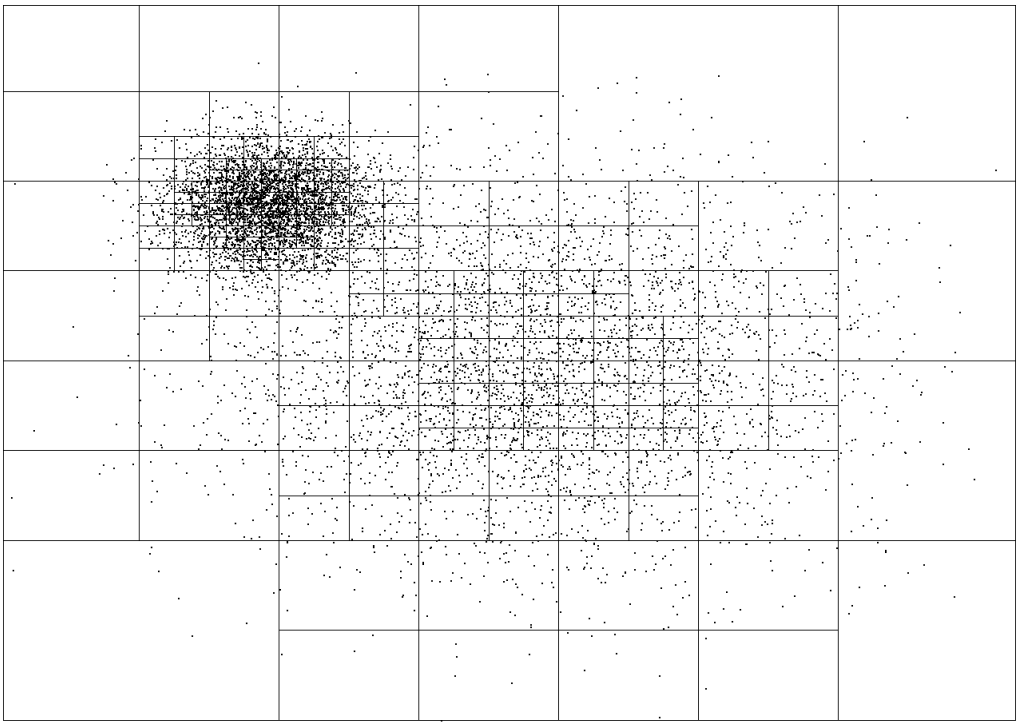
Entretanto, modelagens baseadas em ACs freqüentemente utilizam variações como as mencionadas na Seção 2.1.5.

Os ACs para os quais tanto os nós como as conexões permanecem fixas ao longo do tempo são chamados de *estáticos*. Nas aplicações de ACs à modelagem de sistemas biológicos é freqüente retirar essa restrição para considerar o caso em que tanto os nós "povoados" como as interconexões entre eles mudam ao longo do tempo. Os trabalhos iniciais de LINDENMAYER [46] contêm exemplos deste tipo de AC.

No Capítulo 3, dedicado ao tema Agrupamento de Dados, consideramos um outro tipo de suporte para as células que generaliza o conceito de reticulado. Nesse caso, as células não estão regularmente distribuídas numa malha; entretanto, a posição relativa das células é fixa ao longo do tempo, e não há criação ou destruição de células. O *suporte celular* é construído a partir da divisão progressiva e simultânea em todas as dimensões de um compacto de $\mathbb{R}^n$. O procedimento de divisão sucessiva é análogo ao empregado para gerar as *quad-trees* [24], mas o processo de divisão é controlado baseado em considerações relacionadas com a densidade de pontos na célula. Na Figura 2.2 representamos o suporte gerado a partir de duas nuvens de pontos com diferentes dispersões. No Capítulo 3 é tratado com detalhe o procedimento de geração do suporte celular.



(a) Suporte sem as nuvens de pontos



(b) Suporte com as nuvens de pontos

Figura 2.2: Exemplo de suporte celular

2.2.3 Variantes da vizinhança

Na Seção 2.1.3 foram introduzidas as duas vizinhanças mais comuns na literatura. No entanto, há ACs que consideram duas vizinhanças para uma célula *c* dada: a primeira

vizinhança, chamada *vizinhança de entrada* é formada pelo conjunto de células cujos estados são levados em conta pela regra local para atualizar o estado de c; a segunda, chamada *vizinhança de saída*, especifica o conjunto de células para as quais o estado de c é acessível na hora de atualizar os seus estados. Se o tamanho de as ambas vizinhanças coincide, então o AC é *balanceado* ou *equilibrado.* Os ACs que apresentaremos nos dois capítulos a seguir entram nesta categoria.

Na Seção 2.1.3 indicamos a maneira que a vizinhança de MOORE pode ser generalizada para incluir células além daquelas mais próximas; para isto foi incluído o parâmetro do raio da vizinhança. No caso do suporte celular mencionado na seção anterior, cada célula corresponde a um compacto de $\mathbb{R}^n$; nesse contexto consideramos as seguintes definições:

Definição 1: Duas células c_i e c_j são vizinhas se elas possuem no mínimo um ponto comum.

Definição 2: A vizinhança de MOORE de raio 1 de uma célula c é formada pela própria célula e o conjunto de todas as suas vizinhas de acordo com a definição anterior.

Definição 3: A vizinhança de MOORE de raio $r \in \mathbb{N}^*$ de uma célula c é o conjunto de células da vizinhança de raio $r-1$ mais as suas respectivas células vizinhas.

Na Figura 2.3 aparece representada a vizinhança de raio 2 da célula preta; o suporte celular é o mesmo da Figura 2.2.

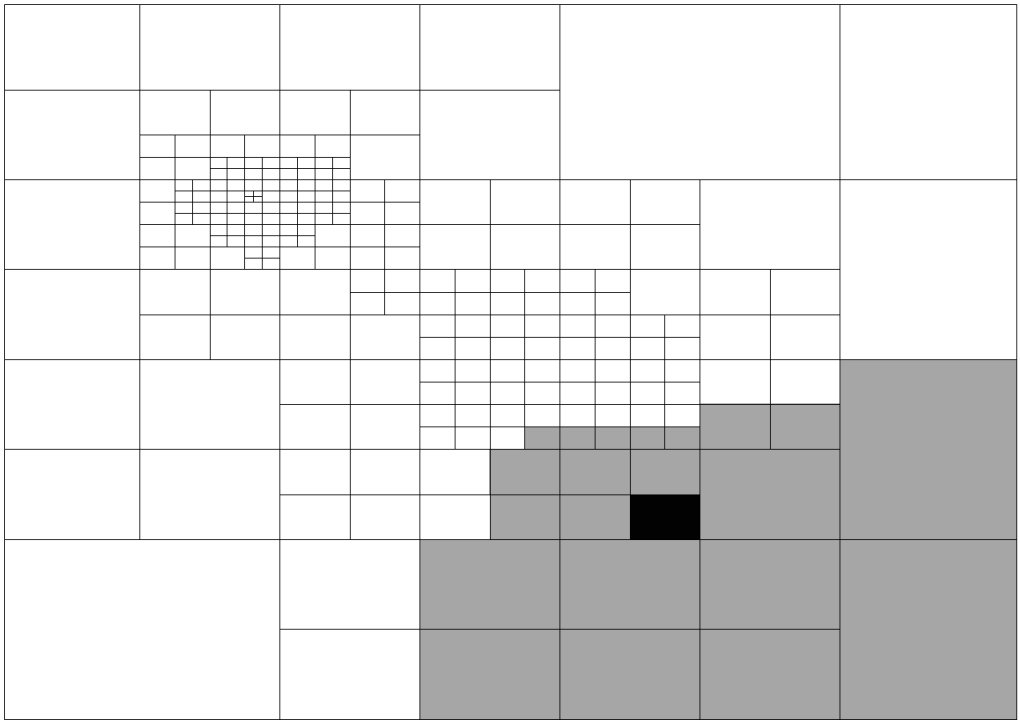


Figura 2.3: Vizinhança de MOORE generalizada de raio 2

No Capítulo 4 é abordado o tema da modelagem de Interações Sociais. A vizinhança considerada nesse contexto é mista: ela tem uma componente fixa criada a partir de uma vizinhança tipo MOORE, e uma segunda componente estocástica, cuja composição é determinada escolhendo aleatoriamente uma quantidade fixa de células[1] em cada passo da simulação. Ainda que o conjunto de células que integram a vizinhança seja diferente em cada passo, a cardinalidade desse conjunto é sempre a mesma para todas as células. A justificação para a escolha desse tipo de vizinhança híbrida é fornecida no Capítulo 4.

2.2.4 Variantes da regra local δ

A definição clássica de autômato celular considera regras de mudança de estado determinísticas. Entretanto, em muitas aplicações, por exemplo TURNER, BEGON e BOWERS [77] estudando a transmissão de patógenos, consideram regras não-determinísticas. Analogamente, no Capítulo 4 consideramos uma regra local estocástica para determinar o próximo estado de uma célula sob influência da vizinhança. Este tipo de autômato celular cuja regra pode mudar no tempo é classificado como *híbrido* [66]. Os ACs híbridos têm sido objeto de desenvolvimentos baseados em VLSI[2] com o nome de autômatos celulares *programáveis.* No caso dos ACs embarcados em VLSI geralmente há uma entrada cujo estado em cada passo determina a escolha de uma regra local dentre várias possíveis.

2.3 Entropia de informação e informação mútua

Nesta seção são apresentados conceitos a serem empregados no Capítulo 3 relativo ao Agrupamento de Dados. Nesse contexto específico, as variáveis consideradas possuem valores discretos; por esse motivo, todas as definições e considerações a seguir estão restritas ao caso de variáveis aleatórias discretas.

Em consideração ao Capítulo 3, tudo o que é preciso considerar são duas variáveis aleatórias discretas, X e Y junto com suas respectivas distribuições de probabilidade. Estas distribuições de probabilidade são estimadas a partir das distribuições de freqüência dos valores das variáveis aleatórias discretas. Em particular, interessa avaliar o grau de dependência entre as duas variáveis aleatórias. Para o estudo desta dependência dispomos somente das distribuições de freqüência e de diversas medidas de independência estatística. A abordagem escolhida se baseia no conceito de *entropia da informação* introduzido por C. SHANNON em 1948 [68], conceito este estruturalmente semelhante ao conceito de entropia usado por BOLTZMANN no contexto da mecânica estatística. O motivo para esta escolha encontra-se, como veremos no próximo capítulo, na proximidade conceitual entre o processo de construir o suporte de um autômato celular e o processo de codificar uma

[1] As células escolhidas aleatoriamente não podem pertencer à componente fixa.
[2] *Very Large Scale Integration*

nuvem de pontos representantes do conteúdo de um banco de dados.

As definições a seguir encontram-se encadeadas, um conceito baseado no anterior. O processo de construção do conceito de *informação mútua*, nosso destino final, somente requer noções elementares de probabilidade.

2.3.1 Auto-informação

Seja X uma variável aleatória e x um possível valor que ela pode assumir; a auto-informação é uma medida da informação contida no fato de que $X = x$, onde esta expressão denota que a variável aleatória X assumiu o valor x. Segundo SHANNON, a auto-informação deve depender inversamente da probabilidade da variável assumir um valor específico; assim fatos pouco prováveis são mais significativos, e por esse motivo eles são portadores de mais informação. Por outro lado, se X pode assumir independentemente tanto o valor x quanto o valor y, então a auto-informação contida no fato de que X assuma uma composição de ambos os valores deve ser igual à soma das auto-informações correspondentes a x e y. Dito de outro modo, a auto-informação é aditiva com respeito à composição de eventos independentes.

A expressão analítica da auto-informação associada ao fato que $X = x$ é:

$$I(X = x) = \log_2(\frac{1}{P(X = x)}) \tag{2.1}$$

onde $P(X = x)$ representa a probabilidade de que a variável aleatória X assuma o valor x. Usualmente é empregado o logaritmo de base 2, resultando ser o *bit* a unidade de medida para a auto-informação[3].

2.3.2 Entropia de informação[4]

SHANNON desejava caraterizar uma fonte de informação do ponto de vista da quantidade de informação que ela pode produzir. Para isso, ele modelou a fonte de informação como uma variável aleatória[5] X, em principio discreta. Essa variável aleatória pode assumir valores dentre um conjunto finito numerável de símbolos. Cada símbolo tem associada uma probabilidade de ser produzido -ou gerado- pela fonte. Destarte, a fonte fica caraterizada pelo conjunto de símbolos (ou valores discretos) -chamado de *alfabeto*- que ela pode gerar e pela distribuição de probabilidade de produção dos símbolos. Na seção anterior foi apresentada a auto-informação como a informação contida no fato da variável aleatória assumir um valor particular. A forma que SHANNON achou de caraterizar a

[3]Outras unidades de medida podem ser empregadas desde que a escolha da base do logaritmo seja diferente. O *nat* é a unidade resultante de usar logaritmo neperiano.

[4]Também conhecida como entropia de Shannon.

[5]Mas precisamente, ele modela a fonte de informação como um processo Markoviano.

fonte de informação foi usar o *valor esperado* da auto-informação da variável aleatória associada. Ele chamou a este valor esperado de *entropia de informação* da fonte, devido a sua semelhança com a expressão de entropia introduzida por BOLTZMANN no seu famoso teorema H da mecânica estatística. Portanto, se designamos por $P(X)$ à distribuição de probabilidade de ocorrência de cada valor que a variável aleatória X pode assumir, resulta que a entropia de informação de X pode ser calculada usando:

$$H(X) = E[I(X)] = -\sum_{i=1}^{n} P(X = x_i) \log_2(P(X = x_i)) \tag{2.2}$$

2.3.3 Entropia condicional

SHANNON introduz o conceito de *entropia condicional* ao considerar duas variáveis aleatórias, X e Y, ambas modelos de uma mesma fonte de informação. Consideremos o problema de avaliar a quantidade de informação em comum que elas têm. Seja $P(X = x, Y = y)$ a probabilidade conjunta de que a variável X assuma o valor x e, simultaneamente, a variável Y assuma o valor y. A entropia da conjunção de X e Y resulta de generalizar a Equação 2.2:

$$H(X,Y) = -\sum_{i,j} P(X = x_i, Y = y_j) \log_2(P(X = x_i, Y = y_j)) \tag{2.3}$$

Levando em conta que

$$H(X) = -\sum_{i,j} P(X = x_i, Y = y_j) \log_2(\sum_{j} P(X = x_i, Y = y_j))$$

resulta que

$$H(X,Y) \leq H(X) + H(Y) \tag{2.4}$$

Consideremos agora a probabilidade condicional $P(Y = y|X = x)$, que expressa a probabilidade da variável Y assumir o valor y dado que a variável X tem o valor x:

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} \tag{2.5}$$

Suponhamos que a variável aleatória X tem o valor x; nesse caso, a entropia da variável Y pode ser calculada adaptando a Equação 2.2:

$$H(Y|X = x) = -\sum_{j} P(Y = y_j|X = x) \log_2(P(Y = y_j|X = x)) \tag{2.6}$$

Shannon define a entropia condicional de Y dada X como a média da entropia de Y

dada por 2.6 ponderada pela probabilidade de X assumir o valor particular x. Ou seja:

$$\begin{aligned} H(Y|X) &= -\sum_{i,j} P(X=x_i)P(Y=y_j|X=x_i)\log_2(P(Y=y_j|X=x_i)) \\ &= -\sum_{i,j} P(X=x_i,Y=y_j)\log_2(\frac{P(Y=y_j,X=x_i)}{P(X=x_i)}) \\ &= -\sum_{i,j} P(X=x_i,Y=y_j)\log_2(P(X=x_i,Y=y_j)) + \sum_{i,j} P(X=x_i,Y=y_j)\log_2(P(X=x_i)) \\ &= H(X,Y) + \sum_{i} P(X=x_i)\log_2(P(X=x_i)) \\ &= H(X,Y) - H(X) \end{aligned} \tag{2.7}$$

Podemos reescrever 2.3 usando 2.7 de modo geral:

$$\begin{aligned} H(X,Y) &= H(Y|X) + H(X) \\ &= H(X|Y) + H(Y) \end{aligned} \tag{2.8}$$

As Equações 2.8 facilitam a interpretação da definição de entropia condicional; assim, a entropia conjunta, ou seja a informação que temos ao considerar simultaneamente as duas variáveis aleatórias, é composta pela entropia numa das variáveis mais a entropia condicional da outra. Deduzimos que a entropia condicional de uma variável representa a informação que não está presente na outra variável. Destarte, combinando 2.4 e 2.8 obtemos $H(Y|X) \leq H(Y)$; a igualdade é válida no caso que as duas variáveis sejam independentes.

2.3.4 Informação mútua

Dadas duas variáveis aleatórias X e Y , a *informação mútua* $I(X,Y)$ avalia a quantidade de informação que ambas variáveis têm em comum. Levando em consideração que a entropia condicional $H(Y|X)$ representa a informação contida na variável Y mas que não está presente na variável X, deduzimos que

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned} \tag{2.9}$$

Se combinamos a definição anterior com a Equação 2.8, resulta:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2.10}$$

Destarte, o cálculo da informação mútua de duas variáveis pode ser feito a partir das entropias individuais e da entropia conjunta. Estas podem ser estimadas empregando as

distribuições de freqüência das variáveis, assim como a distribuição de freqüência conjunta. Esta abordagem é a empregada no Capítulo 3 na construção de um autômato celular.

2.3.5 Divergência de KULLBACK-LEIBLER

Uma abordagem alternativa para calcular a informação mútua de duas variáveis aleatórias se baseia no conceito da *divergência* de KULLBACK-LEIBLER [43]. Sejam X e Y duas variáveis aleatórias, $P(X)$, $P(Y)$ e $P(X,Y)$ as distribuições de probabilidade associadas. As duas situações extremas vinculando as duas variáveis são:

- As duas variáveis são independentes, então $P(X,Y) = P(X)P(Y)$
- Uma das variáveis está totalmente determinada pela outra, então $P(X,Y) = P(X) = P(Y)$

No caso geral, situado entre esses dois extremos, temos que $P(X,Y) \neq P(X)P(Y)$. Uma maneira de avaliar o grau de dependência entre as variáveis é calculando a diferença ou divergência que há entre a distribuição conjunta, conhecida ou estimada por algum meio, e a distribuição conjunta esperada para o caso em que as duas variáveis são independentes. Ou seja, o grau de dependência se reflete na divergência entre $P(X,Y)$ e a distribuição obtida a partir do produto das distribuições individuais, $P(X)P(Y)$. A divergência de KULLBACK-LEIBLER é uma medida geral da divergência entre duas distribuições de probabilidade. Sejam P_1 e P_2 duas distribuições de probabilidade; designamos com $J(P_1,P_2)$ a medida de quanto diverge P_2 de P_1, e sua expressão é a seguinte:

$$J(P_1,P_2) = \sum_x P_1(x)\log_2(\frac{P_1(x)}{P_2(x)}) \tag{2.11}$$

Calculemos[6] a divergência de KULLBACK-LEIBLER de $P(X,Y)$ em relação a $P(X)P(Y)$:

$$\begin{aligned} J(P(X,Y),P(X)P(Y)) &= \sum_{i,j} P(x_i,y_j)\log_2(\frac{P(x_i,y_j)}{P(x_i)P(y_j)}) \\ &= \sum_{i,j} P(x_i,y_j)\log_2(P(x_i,y_j)) - \sum_{i,j} P(x_i,y_j)\log_2(P(x_i)) \\ &\quad - \sum_{i,j} P(x_i,y_j)\log_2(P(y_j)) \\ &= -H(X,Y) + H(X) + H(Y) \\ &= I(X,Y) \end{aligned} \tag{2.12}$$

A Equação 2.12 estabelece que a informação mútua entre duas variáveis aleatórias, tal como foi definida na Equação 2.9, representa a divergência entre a distribuição de

[6] Em benefício da compacidade, denotamos $P(X = x_i)$ como $P(x_i)$

probabilidade conjunta e o produto das distribuições de cada variável. Quanto maior a divergência, maior a dependência entre as variáveis e maior a informação que elas possuem em comum.

2.4 Algoritmos de detecção de agrupamentos de dados

O conjunto de técnicas para a detecção de agrupamentos de dados (*clustering*) faz parte da área de mineração de dados (*datamining*). Detectar (ou criar) agrupamentos num conjunto de dados consiste em separá-los em subconjuntos homogêneos utilizando algum critério. Esta definição se desdobra em muitos casos particulares, dependendo da natureza do conjunto de dados e do critério empregado para estabelecer a composição de cada subconjunto (ou *cluster*). Diversas taxonomias têm sido criadas para classificar a pletora de algoritmos de *clustering*. Cada taxonomia usa um critério diferente para construir a classificação [33]; os mais conhecidos entre eles são:

- a representação dos dados de entrada.
- a representação da saída gerada pelo algoritmo.
- o tipo de distribuição de probabilidade (também chamado de modelo) que descreve as freqüências de aparição dos dados; este caso já implica em uma restrição nos algoritmos a serem considerados.
- a função de custo a ser otimizada, sendo que ela pode ou não ser explícita.
- o sentido usado na construção dos agrupamentos, podendo este ser aglomerativo ou divisivo (também conhecidas como abordagens *bottom-up* ou *top-down*).

Os dados a serem agrupados em geral possuem uma origem comum, o que permite uma representação uniforme. Caso essa representação uniforme não esteja disponível para os "dados brutos", é preciso considerar como mínimo uma etapa de pré-processamento. Exemplos de conjuntos de dados podem ser: séries temporais das leituras de uma rede de sensores, uma coleção de documentos extraídos da Internet, registros de uma tabela de clientes, observações astronômicas, etc. No contexto do presente trabalho, consideraremos somente conjuntos de dados onde cada instância (também chamada *tupla*, registro ou ponto) é descrita por uma lista ordenada de valores numéricos não categóricos. Destarte, vamos supor que cada tupla é composta por d atributos (também chamados dimensões, componentes, caraterística ou campos). Assim, o conjunto de dados com N instâncias forma uma matriz $N \times d$. Uma visão alternativa para o problema de detectar agrupamentos é considerar a escolha de elementos do conjunto potência do conjunto de todos os dados. A forma como esta escolha é feita caracteriza o algoritmo.

A maneira com que os *clusters* são criados tem diversas abordagens, cada uma das quais implica numa série de hipóteses, sejam estas explícitas ou implícitas, em relação aos dados. A seguir apresentaremos as principais estratégias básicas, assim como as combinações mais recentes. As referências [10, 32, 34, 60] apresentam revisões dos algoritmos de *clustering* mais conhecidos. A referência [27] apresenta uma taxonomia de algoritmos de *clustering* desde o ponto de vista da programação matemática.

2.4.1 Algoritmos hierárquicos

Nos chamados algoritmos hierárquicos é criada uma árvore, onde cada nó é um elemento do conjunto potência do conjunto contendo todos os dados. Todos os nós com uma altura determinada dentro da árvore formam uma partição ou classificação possível. Todos os nós representam subconjuntos disjuntos de dados. A forma como esses nós (i.e. subconjuntos) são escolhidos depende da estratégia do algoritmo hierárquico. As estratégias que começam pelas folhas da árvore, subconjuntos unitários, são os algoritmos hierárquicos *aglomerativos* ou *bottom-up*. Exemplos deste tipo de algoritmos são os *single linkage*, *complete linkage* e *average linkage*. Em cada caso, um critério diferente é escolhido para decidir se dois nós da árvore são filhos de um terceiro. Dito de outro modo, dependendo do critério empregado, dois subconjuntos são substituídos pelo conjunto união; este processo é repetido até chegar ao conjunto de todos os dados.

Consideremos o algoritmo *single linkage* e a distância Euclideana. Aqueles pares de folhas mais próximas são fusionadas gerando os nós do nível seguinte da árvore. Por este motivo o algoritmo também é conhecido como “do vizinho mais próximo” (*nearest neighbor*). Para subconjuntos com mais de um elemento, a distância entre eles é aquela que há entre os dois elementos mais próximos de cada subconjunto. Esta forma de medir a distância entre subconjuntos tem a desvantagem de fusionar subconjuntos ainda que seus centróides sejam distantes. Por este motivo outras abordagens definem de outro modo a distância entre nós da árvore. Por exemplo, no caso da *average linkage* a distância entre subconjuntos é aquela que existe entre os centróides.

Em geral estes algoritmos hierárquicos criam a chamada *matriz de distâncias* para decidir a agregação de subconjuntos. Os elementos dessa matriz representam as distâncias entre os nós da arvore. Toda vez que dois nós são fusionados num terceiro, uma coluna e uma linha da matriz são suprimidos.

As diferenças entre as diversas versões de algoritmos hierárquicos está na medida de distância empregada (Euclideana, Manhattan, Mahalanobis, etc.) e o critério usado para decidir se dois subconjuntos devem ser fusionados. Estes algoritmos têm como vantagem poder detectar *clusters* com qualquer distribuição de pontos no seu interior. No entanto, a sua concepção não os faz naturalmente paralelizáveis pois devem considerar combinações de pontos para avaliar as distâncias. Outra limitação está no fato de que todos eles

precisam de um critério alheio ao algoritmo para decidir a altura dos nós que representam os subconjuntos que compõem a classificação almejada.

Os algoritmos hierárquicos têm sido empregados com sucesso combinados com os algoritmos chamados particionadores; mais adiante indicaremos alguns exemplos conhecidos.

2.4.2 Algoritmos baseados em modelos estatísticos

Este tipo de algoritmo supõe que a distribuição dos dados pode ser modelada por uma combinação de distribuições de probabilidade. Esta combinação de distribuições de probabilidade é conhecida como mistura finita de modelos (*finite mixture models*). Nesta abordagem, cada ponto da nuvem de dados tem associada uma probabilidade de ter sido gerado a partir de cada uma das distribuições que compõem a mistura. Destarte, a classificação de um ponto como pertencente a um *cluster* é probabilística; esta caraterística libera aos algoritmos baseados em modelos da restrição de considerar somente subconjuntos disjuntos, como no caso dos algoritmos hierárquicos mencionados acima.

O problema de detectar agrupamentos se transforma no problema de achar distribuições de probabilidade em que os dados sejam verossímeis. Como conseqüência, devemos acrescentar, além da hipótese da quantidade de *clusters*, o modelo de distribuição que melhor descreve os dados. Por motivos práticos, este tipo de algoritmo coloca as supostas distribuições de probabilidade no Leito de Procusto das distribuições Gaussianas. Destarte, este tipo de algoritmo de detecção de agrupamentos se propõe achar a média μ e o desvio padrão σ de cada uma das distribuições que descrevem com maior verossimilhança a distribuição dos pontos. O algoritmo geralmente empregado para a estimação dos parâmetros das distribuições é o conhecido *expectation-maximization* (EM) introduzido por DEMPSTER *et al*. [21]. Após escolher valores iniciais aleatórios para os parâmetros (μ_i, σ_i), com $i = 1,..,k$, k o número de *clusters* a ser detectados, o algoritmo itera os seguintes dois passos:

1. empregando as distribuições com o valor atual dos parâmetros, são calculadas as probabilidades associadas a cada ponto da nuvem de dados. Esta é a fase de *expectation*.

2. com as probabilidades calculadas no passo anterior é gerado um novo conjunto de parâmetros para as distribuições. Este novo conjunto de (μ_i, σ_i) é calculado maximizando a verossimilhança; dito de outra forma, com o novo conjunto de parâmetros há uma melhor adequação entre as distribuições de probabilidade e os dados. A verossimilhança é um indicador que mede a adequação de uma distribuição de probabilidade (ou combinação delas) para predizer a forma como um conjunto de dados se distribui. Informalmente, a verossimilhança é interpretada como a

"probabilidade" que um conjunto de dados tenha sido gerado a partir de uma dada distribuição de probabilidade. O indicador verossimilhança não está limitado ao intervalo $[0,1]$, e portanto ele não corresponde a uma medida de probabilidade.

O algoritmo EM detém as iterações quando não há um aumento significativo na verossimilhança global.

Este tipo de algoritmo requer, como já foi dito, que a quantidade de *clusters* seja especificada de antemão. Além disso, a escolha dos parâmetros (μ_i, σ_i) iniciais tem uma influência muito grande no desempenho do algoritmo. Recentemente BLEKAS e LAGARIS [12] propuseram um método para determinar tanto o número de *clusters* quanto os valores iniciais para os parâmetros das distribuições. Eles criaram uma função potencial inspirados na lei de atração gravitacional. Os máximos locais do potencial são os candidatos a serem centros de *cluster*. O método desenvolvido por eles também fornece valores iniciais para os desvios padrão das Gaussianas que compõem a mistura. Em seguida, o algoritmo EM é usado para achar a mistura de distribuições com maior verossimilhança. A proposta implica num alto custo computacional inicial que em alguns casos pode ser compensado pela velocidade de convergência do algoritmo EM. Entretanto, o método proposto por BLEKAS e LAGARIS para detectar *clusters* é inadequado para grandes bancos de dados. Ele também apresenta a limitação de não poder detectar *clusters* com forma diversa, pois as distribuições favorecem as formas elipsoidais. Além disso, a complexidade algorítmica aumenta quadraticamente com o número de atributos que não possuem independência estatística.

2.4.3 Algoritmos particionadores

Talvez o mais conhecido dos algoritmos de formação de agrupamentos seja o *k-means*, publicado por MACQUEEN[7] em 1967 [48]. Este algoritmo entra na categoria dos chamados algoritmos de *particionamento*. A idéia original busca minimizar a soma das distâncias entre pontos que pertencem ao mesmo agrupamento e o seu centro de massa. O algoritmo parte do suposto de que a quantidade de agrupamentos k é conhecida. A versão clássica do algoritmo começa com a escolha de k pontos "sementes", sendo que os agrupamentos resultam de considerar os subconjuntos de pontos da nuvem mais próximos de cada semente. Sejam $\{c_1, .., c_k\}$ os centros dos *clusters*, e $d(.,.)$ uma medida de distância, sendo que geralmente é escolhida a distância Euclideana. O algoritmo busca minimizar a seguinte expressão:

$$D = \sum_{i=1}^{n} [\min_{j=1..k} d(x_i, c_j)]^2 \tag{2.13}$$

[7]Esta referência é a usualmente aceita como ponto inicial para o algoritmo k-means; entretanto a referência [63]aponta para um trabalho anterior de FORGY em 1965 [25].

onde $d(x_i, c_j)$ representa a distância entre um ponto x_i da nuvem e o centro c_j. A expressão de D recebe várias interpretações:

- Se os centros dos agrupamentos são escolhidos como as médias dos pontos que pertencem ao agrupamento, então o objetivo de minimizar D pode ser interpretado como a minimização da variância intra-*cluster*.
- Analogamente, se interpretarmos os centros dos agrupamentos como os centróides, então minimizar D equivale a criar *clusters* com o menor momento de inércia.
- Os centros podem ser escolhidos entre pontos da nuvem de dados. Nesse caso, eles são os medióides do subconjunto que compõe um *cluster*. Neste caso D pode ser visto como um potencial.

Independentemente de como tenha sido feita a escolha inicial dos centros com a consequente classificação dos pontos da nuvem, o algoritmo clássico busca minimizar D iterando nas seguintes duas etapas:

1. Escolher um novo centro para os agrupamentos formados. A escolha pode ser o resultado de calcular a média ou o centróide dos pontos pertencentes a um mesmo *cluster*, ou selecionando o seu centróide.
2. Usar os centros achados no passo anterior para reclassificar todos os pontos da nuvem.

O algoritmo detém as iterações quando no segundo passo não há mais mudanças na classificação dos pontos. Este método faz com que D decresça monotonicamente até atingir um mínimo local. Dependendo da escolha inicial dos centros, o algoritmo pode convergir para diferentes mínimos locais do potencial. Por este motivo, usualmente são feitas várias rodadas do algoritmo, cada vez com sementes diferentes, escolhendo-se aquela partição que tenha o menor valor para D. Esta metodologia baseada em várias tentativas, sem garantia de chegar a uma solução adequada, não é adequada para grandes bancos de dados.

A simplicidade conceitual desde algoritmo justifica seu amplo uso e os numerosos trabalhos que buscam superar as limitações intrínsecas. Por exemplo, em [39] é proposto o algoritmo PAM onde os centróides são substituídos por medióides; entretanto a complexidade algorítmica é grande: $O(k(n-k)^2)$. Há trabalhos, por exemplo em [51], que tentam reduzir esta complexidade com heurísticas para a escolha das sementes baseada numa amostragem dos pontos. Recentemente REDMOND e HENEGHAN [63] propõem um método baseado no fato de que os centros dos agrupamentos possuem uma densidade maior que na periferia. Para explorar este fato, eles empregam uma estrutura chamada *kd-trees*, criada por FINKEL e BENTLEY [24] em 1974; basicamente, as k sementes são escolhidas nos centros geométricos das folhas mais densas. O inconveniente deste método está na

falta de escalabilidade da construção da *kd-tree* com respeito ao número de dimensões. Num outro trabalho recente, ARTHUR e VASSILVITSKII [1] propõem que as sementes sejam escolhidas de modo probabilístico, sendo que a probabilidade de um ponto da nuvem ser escolhido como semente depende do valor que assume a expressão dada pela Equação 2.13 se ele for efetivamente escolhido. Com este método para a escolha das sementes, eles garantem a melhora da acurácia da partição obtida, assim como um ganho na velocidade de convergência nas iterações. No entanto, o processo de escolha de sementes é custoso do ponto de vista computacional, e esse custo pode não estar compensado pela aceleração da convergência.

Outros trabalhos procuram melhorar o desempenho do k-means indicando procedimentos para fazer os cálculos implicados nos passos de cada iteração. Assim, PELLEG e MOORE [61] usam uma estrutura *kd-tree* para armazenar em cada nó da árvore dados estatísticos da região do espaço por ela coberta. Deste modo, esses dados estatísticos podem evitar que sejam recalculadas todas as distâncias dos pontos contidos nessa região.

Uma abordagem diferente do método de iteração para minimizar a Equação 2.13 é a proposta por XAVIER [83]. O problema de detectar os agrupamentos com menor variância interna é transformado num problema de otimização cujo objetivo é achar os k centróides que minimizem o potencial D. O método de suavização hiperbólica aplicado ao potencial e às restrições possibilita o uso de eficientes métodos de otimização para problemas com funções diferenciáveis.

O algoritmo *k-means* foi modificado por BEZDEK *et al.* [11] para criar o algoritmo *fuzzy c-means*. O intuito é generalizar a expressão da função de potencial 2.13 de maneira que a classificação de um ponto não seja única, isto é, cada ponto pode pertencer a vários agrupamentos com um determinado grau de *pertinência*. A expressão do potencial 2.13 se transforma na seguinte:

$$D = \sum_{i=1}^{n} \sum_{j=1}^{k} \mu_j^m(x_i) d(x_i, c_j)^2 \tag{2.14}$$

onde o parâmetro $m > 1$ regula a forma das funções de pertinência e $\mu_j^m(x_i)$ é o valor da pertinência do registro x_i ao agrupamento com centróide c_j. A vantagem deste algoritmo é a sua capacidade de levar em consideração que as fronteiras entre os agrupamentos podem não estar claramente definidas, e destarte, acrescenta a informação sobre a incerteza da classificação de um registro. Dependendo do valor escolhido para m, as fronteiras são mais ou menos definidas, sendo que no caso extremo ($m \to 1$), são reencontrados os resultados do algoritmo *k-means* clássico. As pertinências de um ponto dado aos diferentes agrupamentos devem somar 1, o que resulta numa competição entre os *clusters* pelo domínio sobre os pontos.

2.4.4 Algoritmos híbridos e bio-inspirados

Algoritmos como o *k-means* ou aqueles baseados em misturas de distribuições de probabilidade com verossimilhança máxima, tendem a formar agrupamentos de forma elipsoidal; por este motivo estes algoritmos não conseguem reconhecer padrões de forma arbitrária. Por outro lado, algoritmos do tipo hierárquico capturam agrupamentos de forma arbitrária mas sofrem do inconveniente de não contar com um critério intrínseco para indicar que nível da hierarquia contém os agrupamentos almejados; por este motivo eles precisam de um critério externo para escolher qual partição é a mais adequada. Uma alternativa freqüentemente usada nesses casos é deixar a decisão a um especialista humano. Algoritmos como o CHAMELEON [38] procuram ganhar as vantagens das duas abordagens citadas combinando em duas etapas diferentes, fases de particionamento e de aglomeração hierárquica; como conseqüência, o citado algoritmo consegue reconhecer agrupamentos com formas arbitrárias. No entanto, permanece a necessidade de especificar o valor de parâmetros.

Uma categoria que vem crescendo em quantidade de exemplos é a dos algoritmos de *clustering* bio-inspirados. Eles utilizam conceitos ou metáforas extraídas de outras áreas do conhecimento, freqüentemente relacionadas com a biologia. Por exemplo, as *redes neuronais artificiais* têm sido aplicadas ao problema de aprendizado não-supervisionado. Um caso destacado é o chamado algoritmo SOM (*Self-Organized Map*) criado por KOHONEN [41, 42].

Outros algoritmos bio-inspirados usam algoritmos genéticos (AG) para gerar as classificações. Por exemplo, SARAFIS *et al*. [65] aplicam AG como forma de otimizar a função de potencial do *k-means*. A vantagem desse algoritmo é a capacidade de detectar agrupamentos cuja forma pode ser elíptica.

Como já foi indicado, um dos problemas de que sofrem muitas abordagens para a detecção de agrupamentos é a necessidade de especificar a quantidade de *clusters* a serem identificados. Este requisito pode ser dispensado se é corretamente explorada a capacidade de auto-organização dos sistemas complexos. As sociedades de insetos apresentam uma dinâmica de sistema complexo onde a organização emerge a partir das regras de comportamento dos indivíduos que as compõem. Esta área de estudo, conhecida como *swarm intelligence* [13], tem fornecido uma pletora de algoritmos para o problema de agrupamento [13, 15, 84, 26]. O sistema imunitário dos animais é um outro exemplo de sistema complexo que pode ser usado como metáfora para o desenho de novas soluções [20, 47, 14].

Capítulo 3

Agrupamento de Dados

3.1 Introdução

Apesar de existirem numerosos algoritmos para detecção de agrupamentos, assim como muitos índices de avaliação das partições geradas por estes, nenhum deles terá a sua proposta de agrupamentos aceita se ela não corresponde àquilo percebido visualmente. Nesse sentido, há muitos trabalhos que procuram uma representação bi ou tridimensional dos dados, como, por exemplo, o clássico mapa de SAMMON [64]. Visualmente reconhecemos a existência de um agrupamento devido à presença de gradientes de densidade na distribuição espacial dos dados, independentemente da forma que adotem os agrupamentos. No presente capítulo, apresentamos um novo algoritmo para formação de agrupamentos em grandes bancos de dados, baseado num autômato celular generalizado. O intuito é explorar a capacidade de auto-organização do autômato para que os agrupamentos emerjam guiados por aquelas células que contenham os máximos locais da densidade de pontos. Descrevemos a seguir os principais algoritmos que guardam alguma relação com o nosso trabalho.

WANG *et al.* introduzem o algoritmo chamado STING [79], projetado para a mineração de grandes bancos de dados espaciais (numéricos e bidimensionais). O algoritmo se baseia na construção de uma árvore cujos nós armazenam medidas estatísticas das células que representam. Assim, começando por uma célula inicial que abrange todos os pontos, sucessivas divisões dão origem às células filhas, em cada uma das quais é armazenado, além do número de pontos, o valor da média, desvio padrão e o tipo de distribuição de probabilidade para cada dimensão. O algoritmo possui uma estrutura simples, adaptável para implementações paralelas. A complexidade do algoritmo é $O(N)$ (sendo N a quantidade de registros) para a construção da árvore, e $O(K)$ – sendo K o número de folhas – para a detecção de agrupamentos.

SHEIKHOLESLAMI *et al.* propuseram o algoritmo WaveCluster [69] que emprega ferramentas utilizadas para o tratamento numérico de sinais. O algoritmo se baseia na apli-

cação da transformada de *wavelets* ao conjunto de valores em cada dimensão. Destarte, os dados passam a ter uma representação no espaço de freqüências; as regiões com uma grande concentração uniforme de pontos se correspondem com regiões de baixa freqüência e grande amplitude, enquanto as bordas dos agrupamentos têm associadas faixas de alta freqüência e amplitude reduzida. Para detectar agrupamentos, basta aplicar filtros no espaço de freqüências: com filtros de altas freqüências são retidas somente as freqüências que foram produzidas pelos centros dos *clusters*; aplicando a transformação inversa, é possível deduzir quais são os pontos localizados nos centros dos *clusters*. Escolhendo convenientemente o tipo de filtro a ser empregado, é possível obter agrupamentos com a resolução desejada. O algoritmo é intrinsecamente imune ao ruído. A sua complexidade algorítmica é $O(N)$ para poucas dimensões; entretanto, a complexidade depende exponencialmente da quantidade de atributos. A tradução da representação no espaço de atributos para a representação no espaço de freqüências é computacionalmente custoso; por este motivo o WaveCluster não é viável em aplicações envolvendo grandes bancos de dados.

HINNEBURG e KEIM apresentaram o algoritmo DENCLUE [29, 30], baseado na determinação das funções de densidade de probabilidade a partir dos valores no espaço de atributos. À diferença de outros algoritmos probabilísticos, o DENCLUE não assume um modelo de distribuição de probabilidade *a priori*. A função densidade de probabilidade é estimada a partir da superposição de *kernels*. Em particular, o DENCLUE emprega *kernels* Gaussianos. A localização dos máximos locais de densidade decorre da superposição. Cada ponto do banco de dados usa o gradiente de densidade para escolher um máximo local. Todos os pontos associados ao mesmo máximo local de densidade são classificados como pertencentes ao mesmo agrupamento. Neste algoritmo, é explicitamente empregado o termo influência para denotar o parâmetro que governa a dispersão de cada *kernel*. Recentemente HINNEBURG e GABRIEL [28] apresentaram um aprimoramento do DENCLUE para que cada ponto selecione o seu máximo local de maneira mais eficiente. A superposição de *kernels* para estimar a densidade de probabilidade é também empregada no algoritmo MEAN SHIFT, criado por COMANICIU e MEER [17]; entretanto, sua aplicação está focada no tratamento de imagens.

Os autômatos celulares se apresentam como uma ferramenta atraente para abordar o problema de detecção de agrupamentos; os motivos são os seguintes:

- a informação é processada localmente, portanto a evolução do autômato pode ser implementada em forma paralela. Isto permite um uso eficiente dos *clusters* de processadores, cuja presença tende a ser ubíqua devido a seu relativo baixo custo;
- eles apresentam a capacidade de se auto-organizar, o que reduz a quantidade de hipóteses a serem feitas sobre os padrões a serem descobertos.

JUANICO *et al.* [37] descrevem um mecanismo geral para a formação de agrupamentos;

eles chamam esse mecanismo de *alelomimese*, que representa a tendência de um indivíduo a se comportar como os seus vizinhos. Os autores utilizam um autômato celular para mostrar que esse mecanismo está na base da emergência de agrupamentos. A aparição dos *clusters* decorre da influência que a vizinhança tem sobre um indivíduo. A alelomimese é freqüentemente encontrada em sociedades de insetos. As formas de organização destas sociedades é fonte de inspiração para a área chamada *swarm intelligence*. XU *et al*. [84] apresentam um algoritmo de *clustering* baseado num autômato celular. Cada ponto do banco de dados é associado a uma "formiga virtual", representada por um agente. Os agentes emulam o comportamento das formigas escolhendo o local para dormir. A proposta de XU *et al.* requer que muitos parâmetros sejam calibrados *ad hoc* a fim de obter resultados satisfatórios.

SHUAI *et al.* [70] introduzem um algoritmo de *clustering* baseado num autômato celular generalizado[1]. Os autores destacam de sua proposta a possilidade de uma implementação paralela e a capacidade de auto-organização. Por estes motivos, SHUAI *et al*. recomendam o seu algoritmo para as aplicações envolvendo grandes bancos de dados. O algoritmo mapeia aleatoriamente os dados a serem agrupados num autômato bidimensional. O estado de cada célula é representado pelo objeto que fica mapeado nela, ou por $\emptyset$ se ela fica vazia. Os dados são realocados estocasticamente até chegar numa distribuição de probabilidade estacionária. O algoritmo requer a escolha de valores para uma grande quantidade de parâmetros: dois por cada célula, mais um parâmetro global T correspondente à temperatura global do sistema. O valor de T determina a taxa de mudanças aleatória que acontecerão em cada passo. SHUAI *et al*. não fornecem no seu trabalho os critérios para a escolha dos valores para esses parâmetros.

O algoritmo, cujo desenho descrevemos neste trabalho, utiliza a evolução de um autômato celular generalizado para inferir da sua configuração global os agrupamentos de dados. Para construir o autômato, a região do espaço ocupada pelos pontos é dividida em sub-regiões, cada uma das quais é representada por uma célula do autômato. A divisão do espaço é feita de modo que o tamanho de uma célula seja inversamente proporcional à densidade de pontos na sub-região associada. Destarte, os núcleos dos agrupamentos, onde a densidade de pontos é maior, ficam representados por um conjunto de células cujos tamanhos são menores que aqueles das células que representam as regiões periféricas. O conjunto de células assim gerado é o suporte celular (introduzido na Seção 2.2.2) do autômato. A regra local de mudança de estado do autômato faz com que cada célula escolha a sua vizinha mais "influente" (na Seção 3.2.4 definimos o conceito de influência). À conseqüência destas escolhas, as vizinhas escolhidas, que chamamos de atratoras, aumentam a sua influência. A cada passo da evolução do autômato, a vizinhança de cada célula cresce, incorporando vizinhas mais distantes. Com a vizinhança ampliada, as cé-

[1] Essencialmente o mesmo trabalho já tinha sido publicado previamente [71] pelos mesmos autores mas com outro título. O trabalho que aqui referimos não faz nenhuma referência a essa primeira versão.

lulas atratoras podem ganhar ou perder influência, dependendo da existência de outras células atratoras concorrentes. Como veremos, as células atratoras se concentram nos núcleos dos agrupamentos. O objetivo no desenho do autômato é que, como resultado de sua evolução, cada agrupamento a ser detectado acabe sendo representado por uma célula atratora junto com todas as células que a escolheram. O máximo local de densidade do agrupamento está localizado nessa célula atratora.

Na próxima seção, descrevemos em detalhe cada componente do nosso autômato celular. Na Seção 3.2.1 é introduzido o procedimento para a construção do suporte celular; graças a ele, os dados ficam associados às células do autômato. A vizinhança de uma célula, assim como o controle do seu crescimento, são tratados nas Seções 3.2.2 e 3.2.3. As definições do conceito de influência e do estado de cada célula são abordados na Seção 3.2.4. A regra local de mudança de estado é especificada na Seção 3.2.5. Uma vez completada a definição do autômato, na Seção 3.3 indicamos a maneira como ele pode ser empregado para implementar o algoritmo de detecção de agrupamentos. Finalmente, na Seção 3.4 são discutidos os resultados dos experimentos computacionais que visam caraterizar o comportamento do algoritmo.

3.2 O autômato celular

Na Seção 2.1.1 apresentamos a definição de um autômato celular na sua versão mais simples. O autômato que descrevemos a seguir generaliza algumas da componentes do quarteto $\mathscr{A} = (\mathbb{Z}^d, \mathscr{S}, \mathscr{N}, \delta)$.As generalizações são três:

1. as células não se encontram associadas a um reticulado, mas a uma estrutura que chamamos *suporte celular*

2. cada célula, além do seu estado, possui outros atributos que participam na regra local de mudança de estado.

3. a vizinhança de cada célula é diferente para cada célula; isto decorre do uso do suporte celular. O tamanho da vizinhança é incrementado em cada passo da evolução do autômato.

3.2.1 O suporte celular

O procedimento utilizado na construção do suporte celular é análogo àquele empregado na construção de uma *kd-tree* [24], e guarda semelhanças com o usado pelo algoritmo STING [79]. Entretanto, a estrutura de dados final não é uma árvore, mas uma lista onde cada elemento representa uma célula. Cada célula armazena a seguinte informação:

1. seus limites geométricos;

2. uma lista de referências aos pontos contidos por ela;

3. uma lista com referências às células vizinhas.

Na presente versão do autômato celular consideramos somente atributos numéricos contínuos; seja d a quantidade de atributos do banco de dados. A região do espaço contendo todos os pontos do banco de dados é representada pela primeira célula (c_0) a ser incluída na lista. c_0 corresponde à primeira geração de células, que denotamos por $g(c_0) = 0$; com cada divisão celular, a geração das células filhas é incrementada em um. O *peso* de uma célula c_i, denotado por $n(c_i)$, coincide com a quantidade de pontos contidos pela célula c_i; decorre que $n(c_0) = N$, sendo N a quantidade total de registros no banco de dados. Seja $w \in \mathbb{N}^*$ o peso máximo que uma célula c_i pode ter: se $n(c_i) > w$ então a célula é dividida. O Algoritmo 1 é o mecanismo básico para a construção do suporte celular.

Algoritmo 1 Primeira versão do algoritmo de construção do suporte celular.

Inicialize a lista do suporte celular com c_0 (célula contendo todos os pontos do banco de dados)

Para cada célula c_i da lista de suporte faça:

se $n(c_i) > w$ então faça:

retire c_i da lista de suporte
biparticione c_i em cada uma das d dimensões gerando assim 2^d células filhas
insira todas as células filhas na lista de suporte

Para cada $w \leqslant N$, corresponde um suporte celular, ainda que a lista de células possa ser a mesma para uma faixa de valores de w. Fixado w, as regiões menos densas do espaço requerem uma menor quantidade de divisões e portanto, têm associadas as células mais volumosas do suporte. Em contrapartida, nos núcleos dos agrupamentos encontram-se as regiões mais densas, as quais requerem muitas divisões, gerando assim as células de menor volume no suporte. Se ulteriormente decrementarmos o valor do peso máximo, as divisões decorrentes são mais freqüêntes entre as células menores. No caso extremo, com $w = 1$, todas as células, sejam estas periféricas ou nucleares, são divididas.

Consideremos o suporte correspondente a um w dado. Sejam c_{max} e c_{min} duas células do suporte tais que não há nenhuma outra célula com volume maior que c_{max}, nem nenhuma outra célula com volume menor que c_{min}. Definimos a *diversidade geracional* Δg_w do suporte correspondente a w, sendo

$$\Delta g_w = g(c_{min}) - g(c_{max})$$

Visto que as células volumosas correspondem às regiões periféricas, enquanto as células menores são criadas nas regiões dos núcleos, decorre que diversidade geracional representa a diferença entre o grau de fragmentação dos núcleos e o das periferias dos agrupamentos. Do ponto de vista da construção do suporte é significativo considerar esta diferença de fragmentações; o objetivo é conseguir um suporte celular que deixe em evidência, por meio do tamanho das células, as regiões correspondentes aos núcleos dos agrupamentos. Por este motivo, a diferença geracional é o parâmetro utilizado na construção do suporte. O Algoritmo 2 especifica o procedimento para construir um suporte celular com diversidade geracional Δg^*.

Algoritmo 2 Procedimento Criar_Suporte_Celular(*conjunto_de_dados*)

Variáveis

c_0: célula inicial contendo todos os pontos do conjunto de dados;
$\mathbb{S}$: conjunto das células do suporte celular;
w: quantidade máxima de pontos numa célula;
Δg_w: diversidade geracional do suporte celular cujas células têm como máximo w pontos;
Δg^*: diversidade geracional crítica;

Inicialização

c_0 := menor região do espaço contendo *conjunto_de_dados;*
$\mathbb{S} := \{c_0\}$;
$w := N/2$;
$\Delta g_w := 0$;

Ações

enquanto Δg_w não atinja o seu valor crítico Δg^* **faça**
  para cada $c_i \in \mathbb{S}$ **faça**
    se $n(c_i) > w$
      Retire c_i de $\mathbb{S}$;
      Biparticione c_i em cada uma das d dimensões gerando $\{c_j\}$, $j = 1, ..., 2^d$;
      $\mathbb{S} := \mathbb{S} \cup \{c_j\}$;
  $w := w - 1$;
retorne $\mathbb{S}$

O valor w^* cujo suporte correspondente tem diversidade geracional Δg^* depende do banco de dados considerado. Consideremos o banco de dados representado na Figura 3.1. Ele foi gerado a partir da mistura de duas Gaussianas com dispersões diferentes, cada uma delas contribuindo com 4000 pontos.

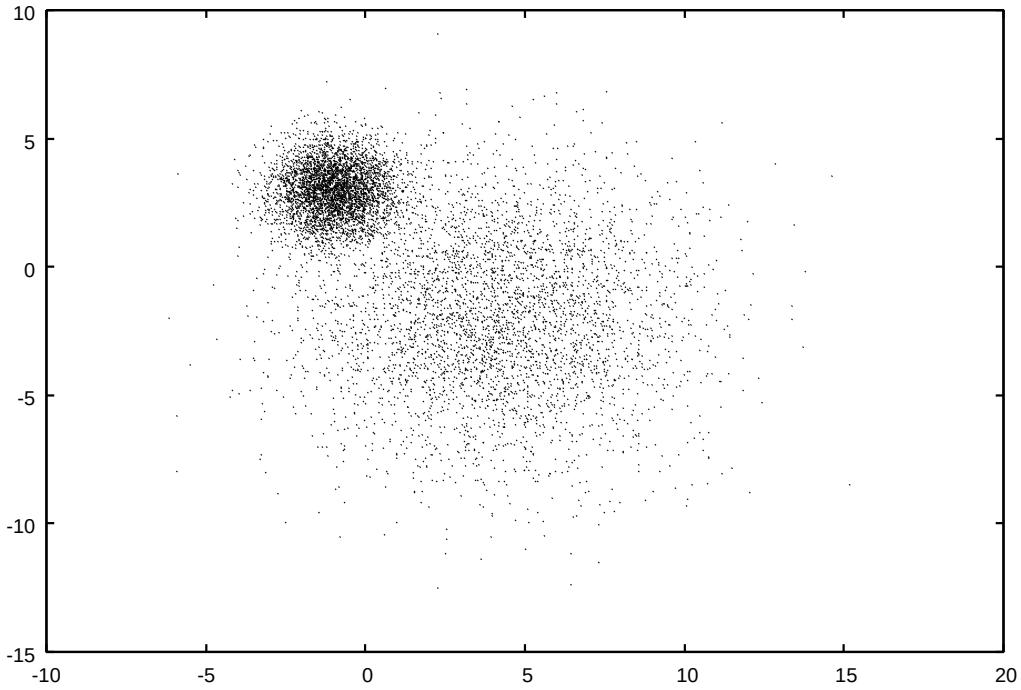


Figura 3.1: Banco de dados com 8000 pontos em duas Gaussianas com diferente dispersão

A Figura 3.2 apresenta os valores de w correspondentes a diferentes Δg, para o banco de dados da Figura 3.1. Para este banco de dados o maior valor possível para Δg é 9: uma redução ulterior do peso máximo acarreta a divisão das células da periferia que implica numa redução da diversidade geracional. Quanto mais uniforme é a distribuição de pontos no bancos de dados, menor é o valor máximo de Δg .

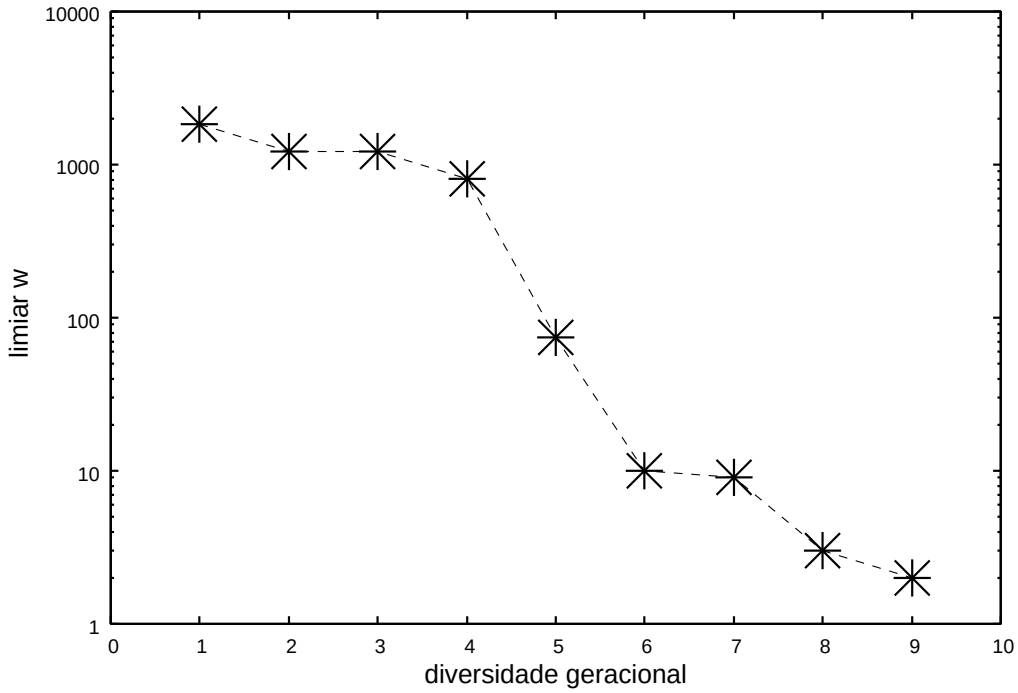


Figura 3.2: Relação entre a diversidade geracional Δg e o peso máximo w correspondente.

A Figura 3.3 mostra o suporte celular para $\Delta g = 5$. O agrupamento da esquerda é mais concentrado, conseqüentemente o seu núcleo está mais fragmentado que o núcleo do agrupamento menos denso.

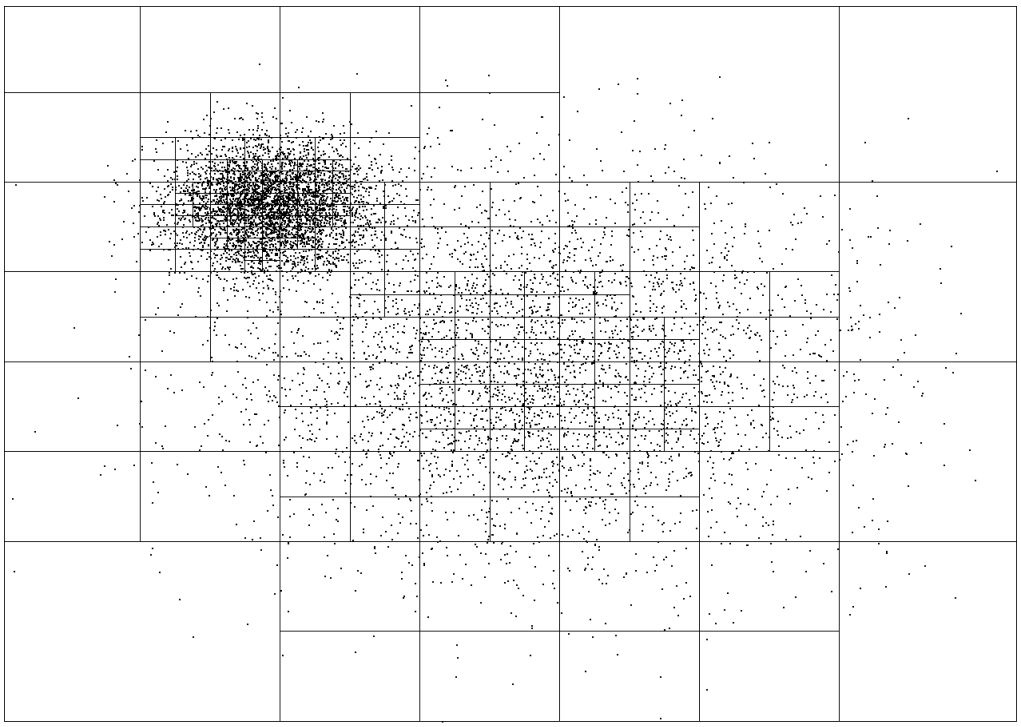


Figura 3.3: Suporte celular no banco de dados de exemplo para $\Delta g = 5$

O suporte celular tem o intuito de deixar em evidência as diferenças de densidade na distribuição de pontos. Os valores da densidade das células representam uma amostragem dos valores da densidade em regiões do espaço. Se a amostragem for adequada, a distribuição de tamanhos das células reflete a distribuição de pontos. A Figura 3.4 mostra o mapa em tons de cinza correspondentes aos valores da densidade em cada célula do suporte. Tons mais escuros correspondem a valores maiores de densidade.

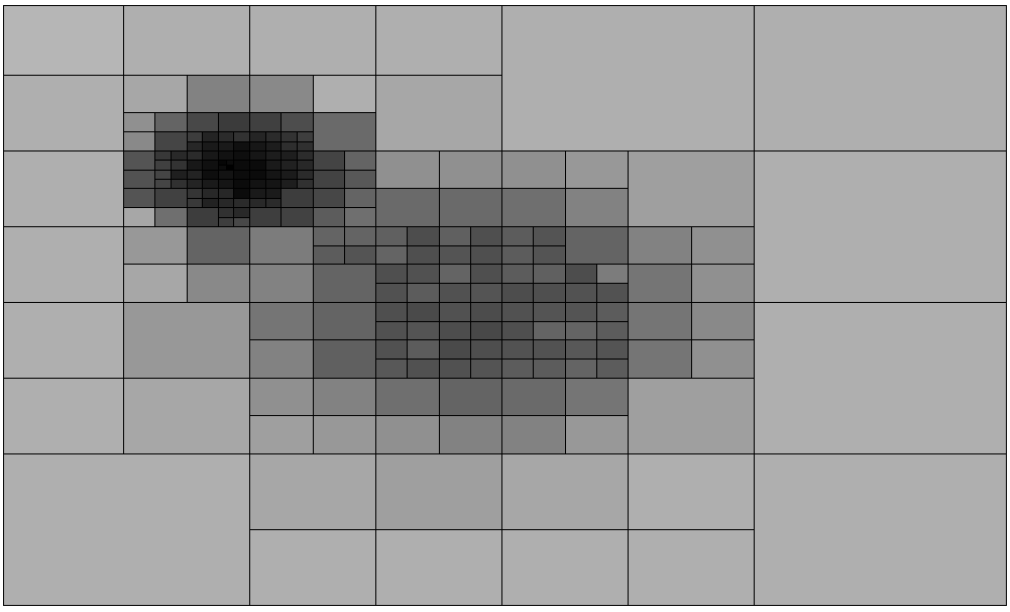


Figura 3.4: Distribuição de densidades no suporte celular do banco de dados do exemplo da Figura 3.1

A diversidade geracional Δg_w de um suporte celular pode ser interpretada como o quociente dos volumes de c_{max} e c_{min}. Equivalentemente, $\Delta g_w + 1$ representa a diversidade de valores para o volume celular. Entretanto, para toda c_i do suporte vale: $0 \leq n(c_i) \leq w$. Decorre que a distribuição de densidade encontra-se discretizada em $(w+1)(\Delta g_w + 1)$ níveis, ou valores, possíveis. Seja V_0 o volume da primeira célula incluída no suporte, e seja g_{max} a geração da célula com o menor volume do suporte associado a Δg_w. O valor máximo possível para a densidade está dado por:

$$D_{max} = \frac{w 2^{dg_{max}}}{V_0}$$

Portanto, a faixa de valores da densidade $[0, D_{max}]$ pode ser discretizada em $(w+1)(\Delta g_w + 1)$ níveis, cada um deles tendo uma largura de:

$$\Delta D = \frac{D_{max}}{(w+1)(\Delta g_w + 1)} \tag{3.1}$$

Independentemente do banco de dados considerado, podemos relacionar a faixa de valores da densidade com a quantidade de níveis de discretização, considerando o *incremento*

relativo de densidade:

$$\Delta D_r = \Delta D \times V_0 \tag{3.2}$$

Este valor é independente das escalas dos atributos, dependendo somente da distribuição de pontos. Assim, por exemplo, levando em consideração os valores da Figura 3.2, construímos a seguinte tabela de incrementos relativos da densidade:

Tabela 3.1: Relação entre a diversidade geracional Δg_w e o incremento relativo da densidade ΔD_r. Para cada valor de Δg_w são fornecidos os valores correspondentes do peso máximo w e da geração mais recente g_{max}.

Δg_w	w	g_{max}	ΔD_r
1	1839	3	0.017
2	1225	4	0.069
3	1225	4	0.052
4	811	5	0.252
5	**74**	**7**	**36.409**
6	10	9	3404.47
7	9	9	3276.8
8	3	10	29127.1
9	2	12	559241

Uma divisão excessiva empobrece a capacidade de detectar diferenças nos valores da densidade, enquanto uma fragmentação pobre não permite capturar com suficiente detalhe as variações da densidade. Há um valor crítico para o incremento relativo, que denotamos ΔD_r^*, para o qual chegamos a uma solução de compromisso entre as duas restrições. Este valor é caracterizado por uma mudança nas ordens de grandeza do incremento relativo. Para o exemplo representado na tabela, o valor crítico ΔD_r^* é atingido para $\Delta g^* = 5$. Em todos os bancos de dados analisados foi possível detectar a existência desse valor crítico, entretanto não dispomos de uma demostração teórica que explique e garanta a sua existência.

Concluindo, a construção do suporte celular consiste em aplicar o Algoritmo 2 até atingir o valor crítico para o incremento relativo de densidade.

3.2.2 A vizinhança $\mathscr{N}$

O autômato celular que estamos considerando usa a generalização da vizinhança de MOORE para o suporte celular. A Figura 2.3, que voltamos a representar na Figura 3.5 para maior clareza, mostra um exemplo de vinhança de MOORE de raio 2 no suporte celular. Inicialmente, todas as células do suporte têm vizinhanças de MOORE de raio 1. A cada passo da evolução do autômato o raio é incrementado, fazendo com que a vizinhança inclua células cada vez mais distantes. Devido ao suporte celular ser finito, o raio tem

um limite superior para o qual a vizinhança de uma célula qualquer abrange o suporte inteiro. No entanto, este caso extremo dissolve a noção de localidade, fundamental para a formação dos agrupamentos, pois eles são essencialmente fenômenos locais. Portanto, é preciso achar um critério para decidir qual é o valor máximo para o raio, compatível com a noção de localidade.

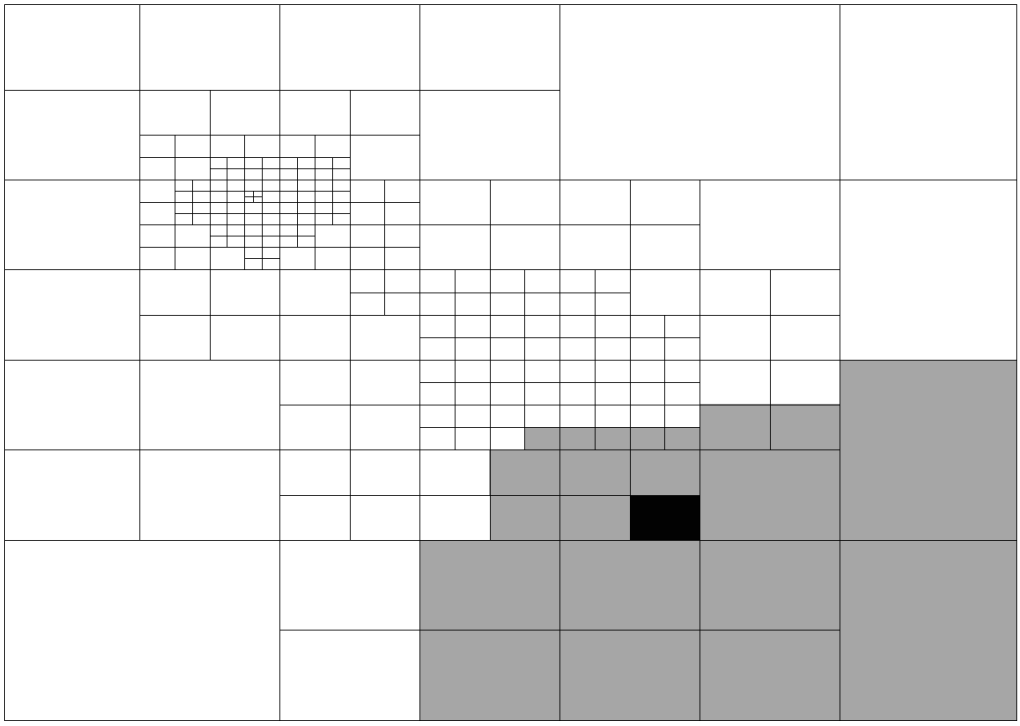


Figura 3.5: Vizinhança de MOORE generalizada de raio 2

3.2.3 Localidade e informação mútua

Consideremos novamente a Figura 3.3. As células do núcleo de um agrupamento possuem aproximadamente o mesmo volume. Definimos o volume de uma vizinhança como a soma dos volumes das células que a compõem. Suponhamos que as células têm suas vizinhança com raio 1. As células do núcleo de um agrupamento também contam com vizinhanças de aproximadamente mesmo volume. Analogamente, as células da periferia de um agrupamento também tendem a ter valores similares tanto para seus volumes como para os volumes de suas respectivas vizinhanças, ainda que diferentes daqueles das células do núcleo. Se consideramos vizinhanças de raio maior que um, a dependência é conservada até que, a partir de um valor do raio, o volume da vizinhança começa a ser independente da célula que estejamos observando. No caso extremo, onde as vizinhanças incluem a maior parte das células do suporte, o volume da vizinhança é o mesmo para todas as células. Ou seja, a dependência entre o volume de uma célula e o de sua vizinhança desaparece à medida em que esta perde sua capacidade de representar uma visão local do entorno da célula. O critério que propomos busca avaliar esta dependência de

volumes como forma de estabelecer se as vizinhanças possuem um tamanho adequado para capturar fenômenos locais. Para isso, utilizamos o conceito de informação mútua introduzido na Seção 2.3.

Seja c_i uma célula do suporte, e $\mathscr{N}(c_i)$ a sua vizinhança; denotamos por $V(c_i)$ o volume da célula, e por $V(\mathscr{N}(c_i))$ o volume da sua vizinhança. A construção do suporte celular é composta por divisões sucessivas da célula inicial; por este motivo, o volume de cada célula é, dependendo da sua respectiva geração, uma fração do volume da célula inicial. Mais concretamente, se $g(c_i)$ é a geração da célula c_i, então temos que:

$$V(c_i) = \frac{V_0}{2^{dg(c_i)}}$$

onde V_0 é o volume da primeira célula incluída no suporte, e d representa a quantidade de atributos no banco de dados. Consideremos as variáveis aleatórias discretas representadas pelos volumes das células e das vizinhanças; suas distribuições de probabilidade dependem do banco de dados em consideração. Como já foi dito no capítulo anterior, a informação mútua $I(V(c_i),V(\mathscr{N}(c_i)))$ avalia a quantidade de informação que ambas as variáveis têm em comum. Substituindo na Equação 2.10 obtemos:

$$I(V(c_i),V(\mathscr{N}(c_i))) = H(V(c_i)) + H(V(\mathscr{N}(c_i))) - H(V(c_i),V(\mathscr{N}(c_i))) \tag{3.3}$$

onde $H(.)$ e $H(.,.)$ denotam a entropia individual e a entropia conjunta respectivamente. Estas podem ser estimadas empregando as distribuições de freqüência das variáveis, assim como a distribuição de freqüência conjunta. Na Figura 3.6 está representada a relação entre a informação mútua e o tamanho da vizinhança para diferentes valores da diversidade geracional no banco de dados do exemplo. Em todos os casos é possível observar que, a partir de um determinado valor do raio, as duas variáveis aleatórias começam a se tornar independentes pois têm cada vez menos informação em comum. A partir desse momento, as vizinhanças respectivas perdem a sua caraterística de representar fenômenos locais. Deduzimos portanto que, a efeito de formar agrupamentos, somente devem ser considerados valores do raio para os quais a informação mútua não decresce. Desde que o raio é incrementado em cada passo da evolução do autômato, este critério permite inferir a quantidade máxima de passos que o autômato deve iterar.

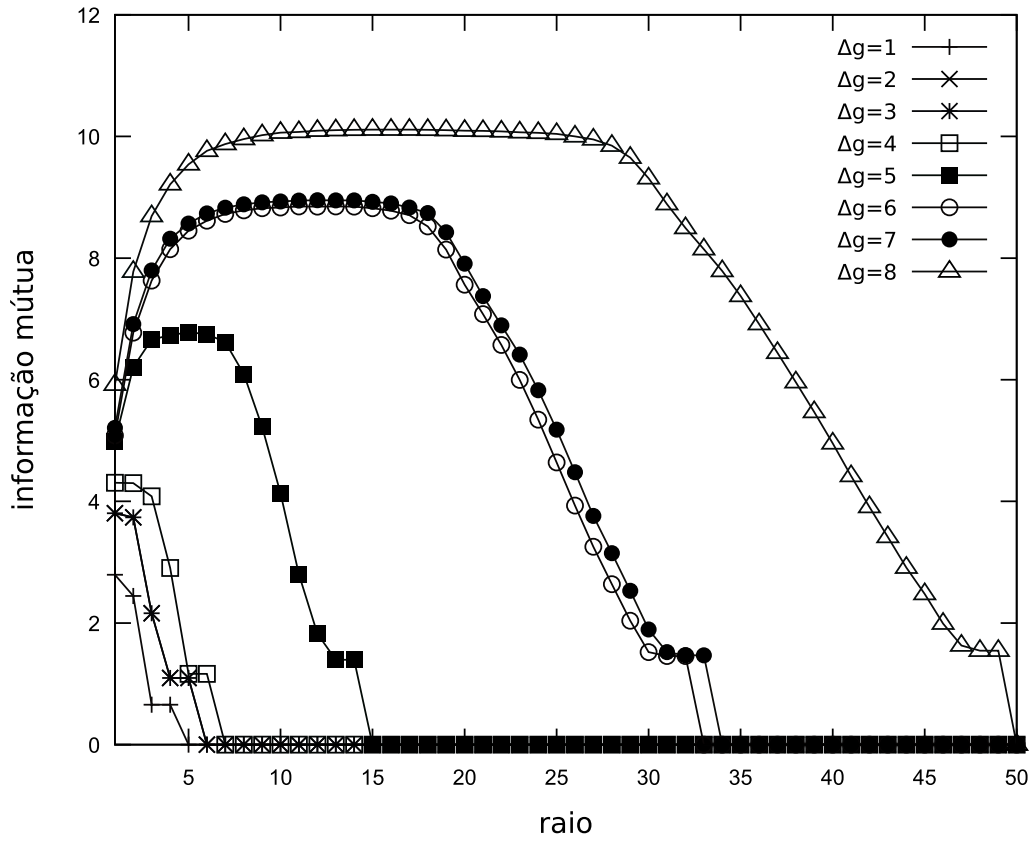


Figura 3.6: Informação mútua para os volumes das células e suas vizinhanças respectivas, para diferentes valores da diversidade geracional, no caso do banco de dados de exemplo.

3.2.4 Atributos das células e o conjunto de estados S

Analogamente à maneira implementada pelo algoritmo DENCLUE, buscamos identificar as regiões contendo os máximos locais da densidade de pontos. Uma primeira tentativa poderia ser simplesmente escolher as células mais densas do suporte; entretanto, esta abordagem possui dois inconvenientes:

- várias células pertencentes ao mesmo agrupamento podem ter o mesmo valor máximo de densidade: qual delas escolher?
- se o banco de dados inclui agrupamentos com concentrações diferentes: como distingüir eficientemente as células nucleares de um agrupamento pouco concentrado das células periféricas de um outro agrupamento muito concentrado?

A estratégia para identificar as células contendo as regiões mais densas dos agrupamentos se baseia no conceito de influência. Buscamos que as células nucleares ganhem influência à medida que o autômato evolui. Para isso, criamos um mecanismo simples para que uma célula possa aumentar a sua influência a partir de suas células vizinhas. Associamos a cada célula c_i um atributo que chamamos *peso acumulado*, que denotamos por $n_A(c_i)$. Inicialmente, o peso acumulado de cada célula tem o mesmo valor que seu própio peso

ou seja, $n_A(c_i) = n(c_i)$. Para cada célula c_i do suporte, definimos a sua *influência* $\mathscr{I}(c_i)$ por:

$$\mathscr{I}(c_i) = D(c_i) \times n_A(c_i) \tag{3.4}$$

onde $D(c_i)$ denota a densidade da célula. Em cada passo da evolução do autômato, cada célula c_i escolhe a *vizinha mais influente*, denotada por $\mathscr{C}(c_i)$. Chamamos *atratoras* àquelas células que escolhem a si mesmas como vizinha mais influente. Toda célula c_i está relacionada com uma célula atratora:

- a relação é direta se $\mathscr{C}(c_i)$ é atratora;
- a relação é indireta se $\mathscr{N}(c_i)$ não contém nenhuma atratora; neste caso, a célula atratora relacionada com c_i é determinada recursivamente escolhendo a atratora relacionada com $\mathscr{C}(c_i)$.

A cada passo da evolução, o peso acumulado de cada atratora é incrementado com o peso das células relacionadas.

O estado de cada célula c_i, que denotamos por $\mathscr{E}(c_i)$, é a referência à atratora com a qual se relaciona. Destarte, o conjunto $\mathscr{S}$ de estados é composto pelo conjunto de referências a cada uma das células do suporte celular.

3.2.5 A regra local de mudança de estado

Denotamos por t o identificador do passo atual do autômato. A cada passo da evolução, cada célula c_i do suporte deve completar duas tarefas:

1. escolher a sua vizinha mais influente, $\mathscr{C}(c_i, t+1)$;
2. definir o seu próximo estado $\mathscr{E}(c_i, t+1)$ determinando qual é a atratora com a qual está relacionada; consecutivamente, acrescentar ao peso acumulado da sua atratora o seu próprio peso $n(c_i)$

Notar que <u>todas</u> as células precisam escolher a sua vizinha mais influente antes de que cada uma possa atualizar o seu estado. Por este motivo, o **Procedimento Atualizar_Estado_Celulas(** $\mathbb{S}$, t **)** – descrito no Algortimo 3 – deve precorrer o suporte celular $\mathbb{S}$ duas vezes: na primeira vez, todas as células escolhem a sua vizinha mais influente; na segunda vez é atualizado o estado de cada célula.

Algoritmo 3 Procedimento Atualizar_Estado_Celulas($\mathbb{S}$, t)

Variáveis

$\mathbb{S}$: conjunto das células do suporte celular;
t: passo atual do autômato celular;
$r(t)$: raio da vizinhança no passo t;
$\mathscr{N}(c_i,t)$: conjunto de células vizinhas da célula c_i;
$\mathscr{C}(c_i,t)$: referência à vizinha mais influente da célula c_i;
$\mathscr{E}(c_i,t)$: referência à atratora relacionada com a célula c_i;
$n(c_i)$: peso da célula c_i;
$n_A(c_i,t)$: peso acumulado da célula c_i;

Ações

para cada $c_i \in \mathbb{S}$ **faça**

$r(t+1) := r(t)+1$;
Atualize $\mathscr{N}(c_i,t+1)$;
Atualize $\mathscr{C}(c_i,t+1)$;
$n_A(c_i,t+1) := 0$;

para cada $c_i \in \mathbb{S}$ **faça**

se $\mathscr{C}(c_i,t+1) = c_i$
$\mathscr{E}(c_i,t+1) := c_i$;
senão
$\mathscr{E}(c_i,t+1) := \mathscr{E}(\mathscr{C}(c_i,t+1),t+1)$;
$n_A(\mathscr{E}(c_i,t+1),t+1) := n_A(\mathscr{E}(c_i,t+1),t+1) + n(c_i)$;

retorne $\mathbb{S}$

A notação $\mathscr{E}(c_i,t+1) := \mathscr{E}(\mathscr{C}(c_i,t+1),t+1)$ deve ser interpretada da seguinte maneira: escolha o valor do próximo estado como sendo igual ao valor do próximo estado da vizinha mais influente. Uma vez que o valor para o próximo estado é a referência à atratora correspondente, a atualização do valor do próximo estado implica numa atualização recursiva de referências.

O Algoritmo 4 deve ser aplicado para simular a evolução do autômato celular.

Algoritmo 4 Simulação do autômato celular.

Variáveis

$\mathbb{S}$: conjunto das células do suporte celular;
t: passo atual do autômato celular;
T_{max}: total de passos de simulação;
$r(t)$: raio da vizinhança no passo t;
$n(c_i)$: peso da célula c_i;
$n_A(c_i,t)$: peso acumulado da célula c_i;

Inicialização

$\mathbb{S} :=$ **Criar_Suporte_Celular(** *conjunto_de_dados*);
$r(0) := 0$;
para cada $c_i \in \mathbb{S}$ **faça**
    $n_A(c_i,0) := n(c_i)$;

Ações

para cada $t := 1,...,T_{max}$ **faça**
    $\mathbb{S} :=$ **Atualizar_Estado_Celulas(** $\mathbb{S}$, t);

3.3 O algoritmo para detecção de agrupamentos

O intuito desta seção é apresentar o nosso algoritmo para detecção de agrupamentos, que doravante chamaremos ABC. Consideremos um banco de dados no qual queremos descobrir os agrupamentos existentes. Para tais efeitos, o algoritmo ABC cria o autômato celular descrito na seção anterior, e o faz evoluir até que as suas células se auto-organizem em agrupamentos. Posteriormente, esses agrupamentos de células são empregados para inferir a classificação dos pontos do banco de dados.

Todo autômato celular pode ser simulado aplicando indefinidamente a regra local para determinar o próximo estado de cada célula. Entretanto, o algoritmo ABC limita a evolução do nosso autômato celular à etapa de emergência dos agrupamentos almejados. A duração desta etapa é determinada usando o critério introduzido em 3.2.3. Cada agrupamento celular é composto por todas as células com o mesmo estado; ou seja, todas as células relacionadas com uma mesma atratora pertencem ao mesmo *cluster*. Uma vez completada a formação dos agrupamentos de células, o algoritmo gera a classificação final dos dados: todos os pontos contidos nas células pertencentes a um agrupamento celular formam um agrupamento de pontos.

Durante a evolução do nosso autômato, a regra local, definida em 3.2.5, é aplicada

a cada célula do suporte. Consideremos a configuração global do autômato no final do primeiro passo. As células atratoras têm o seu respectivo peso acumulado incrementado com o peso de todas as células relacionadas com cada uma delas; as restantes células ficam com peso acumulado nulo. Na primeira metade do segundo passo, cada célula escolhe a sua vizinha mais influente entre as células que compõem a sua nova vizinhança; esta nova geração de vizinhas mais influentes é um subconjunto das atratoras criadas no primeiro passo, pois estas são as únicas a terem influência não-nula. Portanto, as atratoras criadas no final do segundo passo formam um subconjunto das atratoras criadas no primeiro passo. A cada passo da evolução, é criado um conjunto cada vez menor de atratoras. No final da fase de formação dos agrupamentos celulares, somente as atratoras que conseguiram acumular suficiente influência sobrevivem. Elas representam as regiões localizadas nos máximos locais de densidade. O algoritmo ABC explora esta competição entre atratoras, aplicando em cada passo a regra local somente às atratoras. A atualização do estado das restantes células é adiado até o final da fase de formação de agrupamentos. A atualização do estado das células não-atratoras consiste na atualização recursiva das referências às atratoras sobreviventes. O Algoritmo 5 define a seqüência de ações que compõem o algoritmo ABC.

Algoritmo 5 Algoritmo ABC para a detecção de agrupamentos.

Variáveis

conjunto_de_dados: dados a serem agrupados;
$\mathbb{S}$: conjunto das células do suporte celular;
$\mathbb{A}$: subconjunto das células atratoras em $\mathbb{S}$;
t: passo atual do autômato celular;
$r(t)$: raio da vizinhança no passo t;
$\mathscr{C}(c_i,t)$: referência à vizinha mais influente da célula c_i;
$\mathscr{E}(c_i,t)$: referência à atratora relacionada com a célula c_i;
$n(c_i)$: peso da célula c_i;
$n_A(c_i,t)$: peso acumulado da célula c_i ;
$V(c_i)$: volume da célula c_i;
$V(\mathscr{N}(c_i))$: volume da vizinhança da célula c_i;

Inicialização

$\mathbb{S} :=$ Criar_Suporte_Celular(*conjunto_de_dados*) ;
$\mathbb{A} := \emptyset$;
$r(0) := 0$;
para cada $c_i \in \mathbb{S}$ **faça**
    $n_A(c_i,0) := n(c_i)$;

Ações

$\mathbb{S} :=$ **Atualizar_Estado_Celulas(** $\mathbb{S}$, 1);
para cada $c_i \in \mathbb{S}$ **faça**
    se c_i é atratora **então** $\mathbb{A} := \mathbb{A} \cup \{c_i\}$;
enquanto $I(V(c_i),V(\mathscr{N}(c_i)))$ não decresce **faça**
    $t := t+1$;
    $\mathbb{A} :=$**Atualizar_Estado_Celulas(** $\mathbb{A}$, t)
para cada $c_i \in \mathbb{S}$ **faça**
    $\mathscr{E}(c_i,t) := \mathscr{E}(\mathscr{C}(c_i,t),t)$;

Uma interpretação alternativa do algoritmo ABC é a seguinte:

1. a construção do suporte celular equivale à criação de um conjunto de partições, cada uma das quais sendo representada por uma célula do suporte;

2. a formação dos sucessivos conjuntos de atratoras é equivalente à aplicação de um algoritmo aglomerativo hieráquico, no qual um conjunto de atratoras é gerado a partir da aglomeração das atratoras que compõem o conjunto predecessor.

Portanto, o algoritmo ABC pode ser pensado como a composição de uma fase criadora de partições, seguida por uma fase aglomerativa hierárquica. Esta estratégia é a seguida pelo algoritmo CHAMELEON [38], o qual emprega para sua primeira fase uma adaptação do *k-means*, e na segunda fase utiliza variantes do algoritmo *single linkage*. Segundo o estudo empírico de JAIN *et al*. [33], este tipo de algoritmo híbrido apresenta os melhores desempenhos globais.

3.4 Experimentos computacionais

Na presente seção apresentamos um estudo das caraterísticas operacionais do algoritmo ABC. Os experimentos computacionais visam determinar o comportamento do algoritmo com respeito às seguintes caraterísticas:

- sensibilidade ao ruído
- poder de discriminação
- escalabilidade com respeito à quantidade de atributos (dimensões)
- escalabilidade com respeito ao tamanho do banco de dados (quantidade de registros)
- escalabilidade com respeito à quantidade de agrupamentos

O desempenho do algoritmo ABC é comparado com o do algoritmo CHAMELEON [38], empregando-se a implementação disponível no pacote CLUTO[2] [62]. A escolha do CLUTO se baseia nas seguintes caraterísticas:

- ele possui uma longa história de aperfeiçoamento, desde sua apresentação em 1999 até o presente. Sua implementação, feita em linguagem C, se baseia em bibliotecas altamente otimizadas. Seus autores garantem que está livre de *bugs;*
- ele constitui uma referência clássica para a categoria de algoritmos dedicados a grandes bancos de dados com grande número de atributos;
- ele emprega uma abordagem combinada, com fases particionais e fases aglomerativas hierárquicas; segundo o estudo empírico de JAIN *et al*. [33], este tipo de algoritmo híbrido apresenta os melhores desempenhos globais;
- ele pode detectar agrupamentos com formas diversas;

[2]Disponível em http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

- a documentação é abundante e precisa. Após cada execução, o programa fornece um relatório detalhado que inclui uma discriminação do tempo consumido em cada uma das seguintes etapas: entrada e pré-processamento dos dados, a formação dos agrupamentos, e a geração do relatório;
- ele encontra-se disponível para os sistemas operacionais mais difundidos;

CLUTO possui vários parâmetros opcionais para controlar o seu funcionamento. Em todos os experimentos que apresentamos, os parâmetros opcionais mantêm os seus valores por omissão. O único parâmetro requerido por CLUTO, além do nome do arquivo contendo os dados, é a quantidade de agrupamentos a serem detectados. Em todos os experimentos realizados, CLUTO foi executado com o número correspondente de agrupamentos presentes no banco de dados. No contexto de mineração de dados, CLUTO deve ser usado em combinação com algum critério (por exemplo, a de variação de informação [49], ou o índice PBM [59]) para avaliar a qualidade da classificação obtida. A quantidade correta de *clusters* é determinada executando CLUTO com diferentes valores para esse parâmetro, e selecionando a melhor classificação segundo o critério de avaliação utilizado. Portanto, na prática, a obtenção de resultados satisfatórios implica num tempo significativamente maior que aquele consumido por CLUTO para detectar os agrupamentos.

3.4.1 Parâmetros, inicialização e convergência

O algoritmo ABC não possui parâmetros a serem calibrados para mudar o seu comportamento. Os dois parâmetros internos de funcionamento, a diversidade geracional e a quantidade de passos da evolução do autômato celular, têm automaticamente determinados os seus valores durante a execução do algoritmo.

O algoritmo ABC inicialmente considera todos os registros, sem relevância da ordem, para a construção do suporte celular. Não há nenhuma escolha a ser feita na inicialização, como acontece com outros algoritmos que, por exemplo, precisam da seleção de sementes entre os registros do banco de dados.

O algoritmo ABC é determinístico: os agrupamentos detectados são os mesmos, para um banco de dados determinado, toda vez que o algoritmo é executado. Analogamente, a sua condição de término, baseada na variação da informação mútua, é a mesma em todas as execuções. Por este motivo, não corresponde considerar um estudo de convergência ao conjunto de agrupamentos detectados.

3.4.2 Avaliação qualitativa

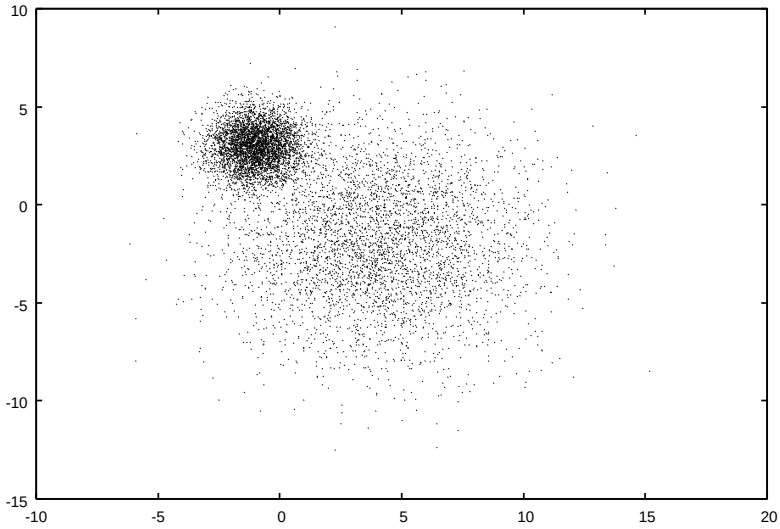
A capacidade do algoritmo ABC para detectar agrupamentos é mostrada usando um conjunto de bancos de dados sintéticos. Em cada um deles, o algoritmo é confrontado com

uma característica no banco de dados cuja presença pode significar uma limitação no desempenho de outros algoritmos. Todos estes bancos de dados são bidimensionais para facilitar sua visualização.

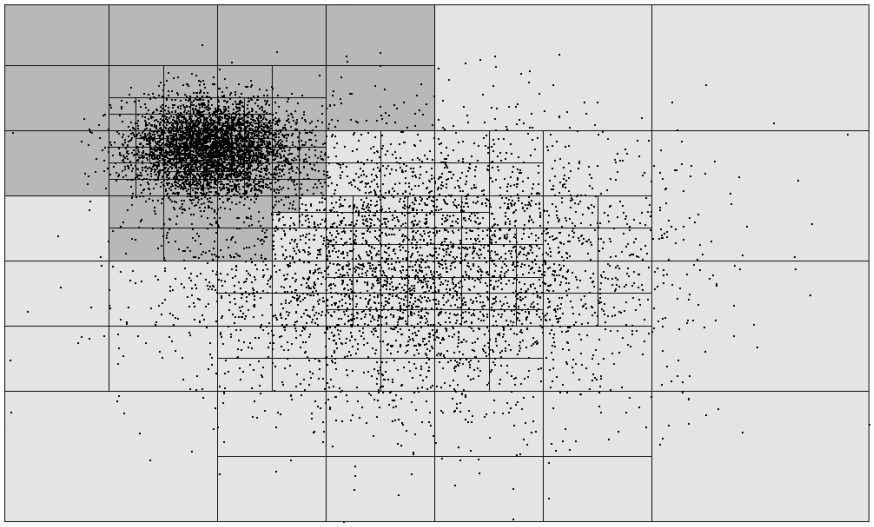
Com o intuito de mostrar o comportamento do algoritmo perante bancos de dados contendo agrupamentos com diversas dispersões, consideramos o banco de dados apresentado na Figura 3.7a. O gráfico corresponde a duas Gaussianas, cada uma delas composta por 4000 pontos, com valores diferentes para o desvio padrão.

Em 3.7b representamos os agrupamentos emergentes de células sobre o suporte celular, após a execução do algoritmo. O agrupamento mais denso possui uma fragmentação maior. Entre as células do seu núcleo encontram-se as atratoras de maior influência do suporte devido a que elas têm a maior densidade. Entretanto, devido à fragmentação, durante a execução do algoritmo as vizinhanças destas atratoras não chegam a incluir células do núcleo do outro agrupamento. Por este motivo, ambos os agrupamentos são detectados sem que a influência de um deles predomine sobre o outro.

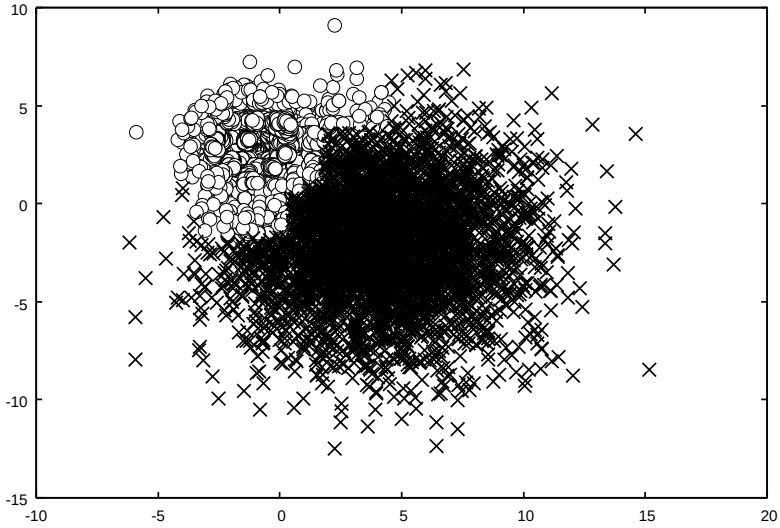
Em 3.7c representamos a classificação final dos pontos. A fronteira entre os dois agrupamentos herda do suporte celular uma forma irregular e abrupta; entretanto, ainda que outros algoritmos possam criar fronteiras mais suaves, a classificação dos pontos limítrofes é intrinsecamente incerta.



(a) Nuvem de pontos original



(b) Agrupamentos de células

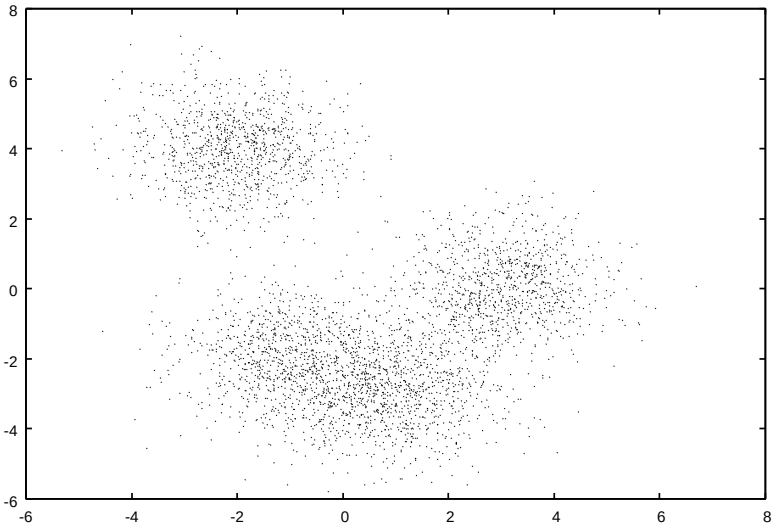


(c) Classificação dos pontos

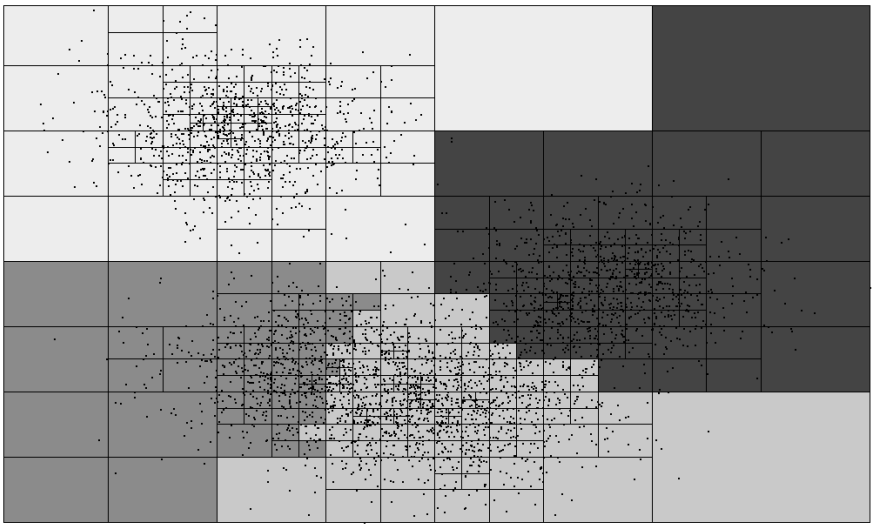
Figura 3.7: Banco de dados composto por duas Gaussianas com diferente desvio padrão, cada uma delas com 4000 pontos.

A capacidade de detectar agrupamentos é condicionada pela capacidade para discriminar *clusters* próximos. A fim de mostrar qualitativamente a capacidade do algoritmo ABC para detectar agrupamentos com diferentes separações, apresentamos na Figura 3.8 um banco de dados composto por quatro Gaussianas com o mesmo desvio padrão, cada uma delas composta por 1000 pontos. Notamos que não é visualmente evidente o reconhecimento dos quatro agrupamentos. Em particular, as duas Gaussianas inferiores apresentam uma superposição significativa.

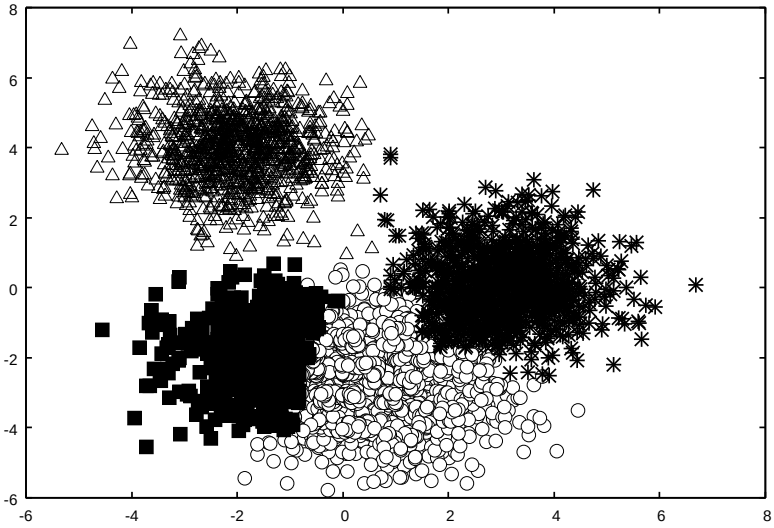
Na Figura 3.8b representamos os agrupamentos de células sobre o suporte celular, após a aplicação do algoritmo. Na região ocupada pelas duas Gaussianas inferiores emergem dois agrupamentos independentes de células. Finalmente, na Figura 3.8c apresentamos a classificação final dos pontos.



(a) Nuvem de pontos original



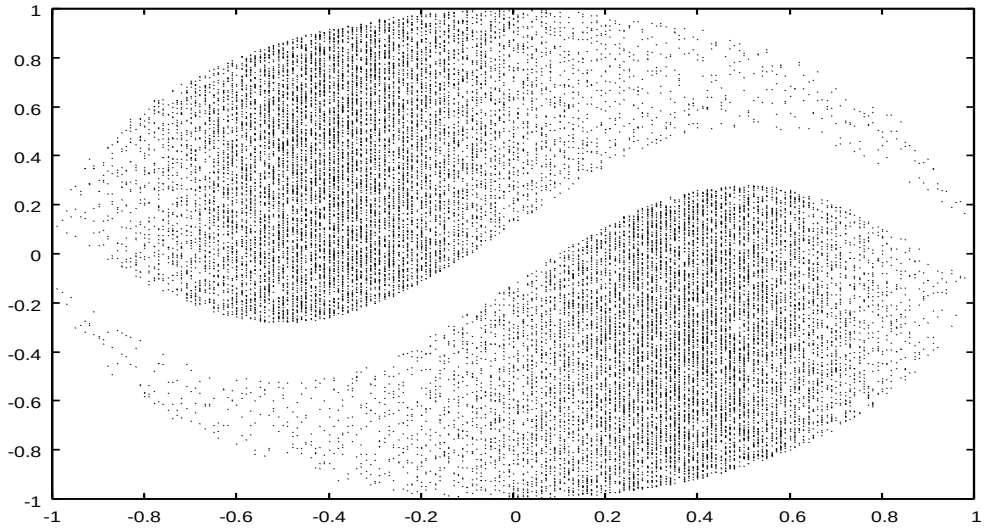
(b) Agrupamentos de células



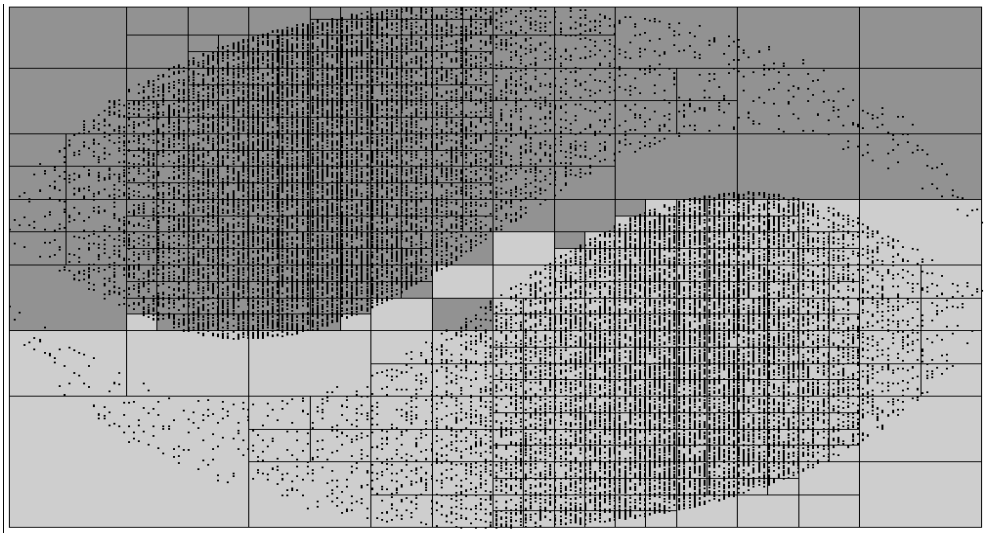
(c) Classificação dos pontos

Figura 3.8: Banco de dados composto por quatro Gaussianas com o mesmo desvio padrão, cada uma delas com 1000 pontos. Os centróides encontram-se a diferentes distâncias entre eles.

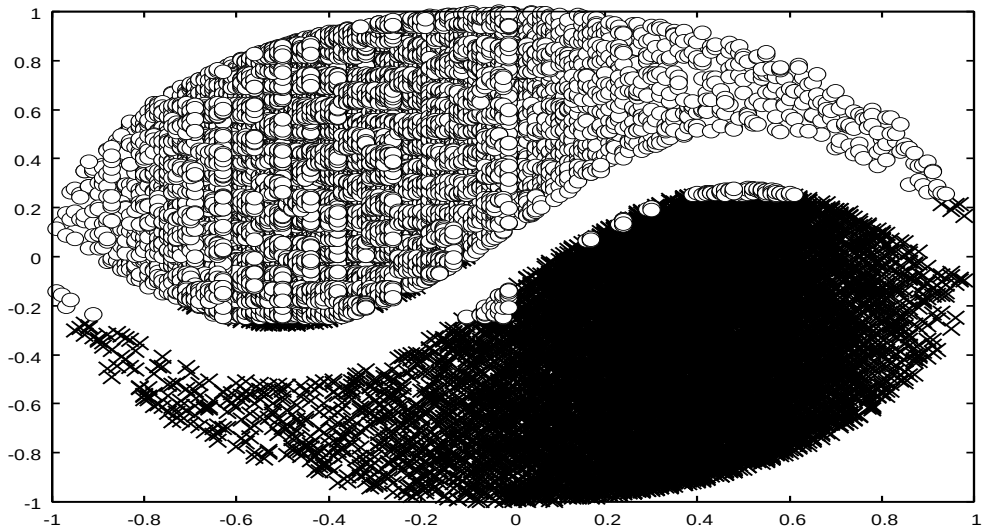
Os algoritmos baseados em partições, por exemplo o *k-means*, têm a limitação de somente reconhecer agrupamentos com forma esférica. Com o intuito de mostrar a capacidade do algoritmo ABC para reconhecer agrupamentos de forma diversa, consideramos o banco de dados representado no gŕafico da Figura 3.9a. Os agrupamentos, cada um deles composto por 12511 pontos, não são linearmente separáveis. Os agrupamentos de células formados após a execução do algoritmo aparecem representados na Figura 3.9b. Os algoritmos cuja função-objetivo implica em minimizar as distâncias *intra-cluster* não são capazes de reconhecer as regiões de baixa densidade como a que separa ambos os agrupamentos; por este motivo, eles classificam os pontos das "caudas" como pertencentes ao mesmo agrupamento que os pontos do núcleo mais próximo. Na Figura 3.9c é fornecida a classificação final dos pontos.



(a) Nuvem de pontos original



(b) Agrupamentos de células



(c) Classificação dos pontos

Figura 3.9: Banco de dados contendo agrupamentos não esféricos, e não linearmente separáveis. Cada cluster é composto por 12511 pontos.

Na Figura 3.10 apresentamos um grafo com a hierarquia de influências para as células que compõem o suporte celular do banco de dados apresentado na Figura 3.9a. Cada nó do grafo representa uma célula; cada aresta conecta a célula na sua origem com a sua vizinha mais influente. A raiz de cada árvore corresponde a uma célula atratora contendo um máximo local de densidade. Devido à forma da distribuição de pontos do banco de dados, as árvores apresentam galhos com grande diversidade de comprimentos. Os galhos mais compridos mostram a hierarquia de influências responsável pelo agrupamento das células localizadas nas “caudas” junto com as células do núcleo correspondente.

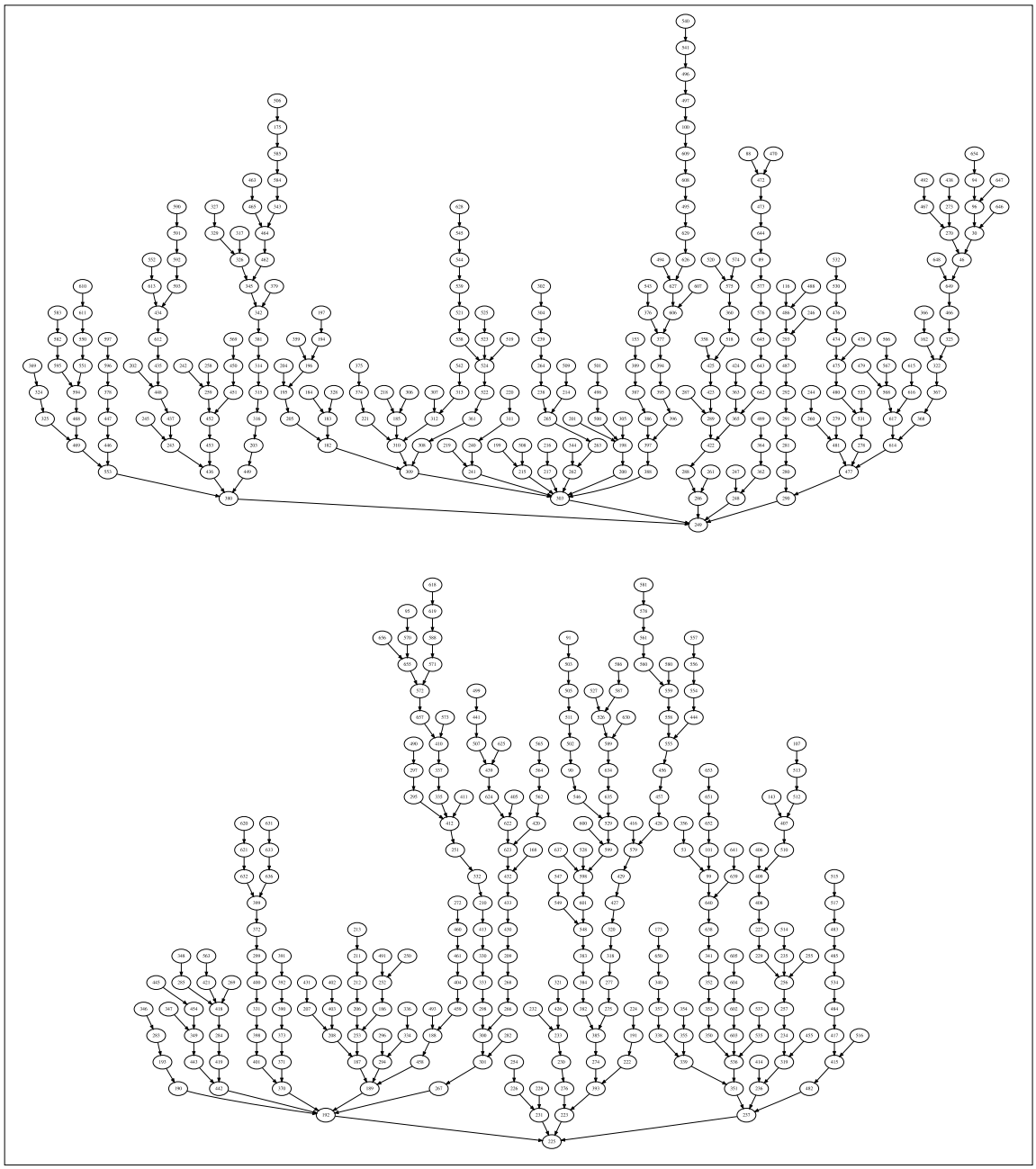


Figura 3.10: Grafo representando a hierarquia de influência das células, para o banco de dados da Figura 3.9.

3.4.3 Sensibilidade ao ruído

No contexto do problema de agrupar dados, o ruído num banco de dados se traduz pela existência de pontos provocados por artefatos, cuja presença pode induzir a uma detecção incorreta de agrupamentos. Os algoritmos de *clustering* que usam a densidade de pontos para a detecção de agrupamentos, como os referenciados na introdução do capítulo, apresentam pouca sensibilidade ao ruído. No caso particular do algoritmo ABC, os pontos são considerados somente na etapa de construção do suporte celular. Este, por sua vez, não considera os pontos individualmente, mas em grupos: o impacto de um dado espúreo é relativo ao tamanho do grupo. A presença de ruído pode afetar a construção do suporte de duas maneiras:

- os pontos espúreos no interior de um agrupamento contam no peso de uma célula; portanto, eles podem implicar que uma célula seja dividida desnecessariamente; entretanto, o impacto desta divisão equivocada sobre o algoritmo é atenuado pelo fato de que o valor da influência leva em consideração tanto a densidade quanto o peso acumulado;
- pontos originados pelo ruído, distantes dos núcleos dos agrupamentos, podem induzir à criação da célula inicial com volume maior; no entanto, as divisões posteriores deixam esses pontos distantes isolados em células de baixa densidade periféricas, sem condições de virar atratoras.

Resulta que a presença de ruído no banco de dados pode implicar na aparição de algumas células desnecessárias, seja no interior dos agrupamentos ou na periferia; entretanto, estas células têm uma capacidade para virar atratoras menor que as respectivas células originais e portanto, o seu impacto é mínimo na deteção dos núcleos dos agrupamentos.

3.4.4 Poder de discriminação

A capacidade do algoritmo ABC para reconhecer agrupamentos independentes é avaliada usando a metodologia empregada por [33]. Para isso, consideramos bancos de dados independentes, cada um contendo duas Gaussianas bidimensionais com desvio padrão unitário; cada um dos agrupamentos é composto por 1000 pontos. Para cada valor de separação entre os centróides – tomando o desvio padrão como referência – foram gerados 10 conjuntos de dados a serem analizados pelo algoritmo ABC. Na Figura 3.11 apresentamos o valor médio da quantidade de agrupamentos detectados para cada separação considerada.

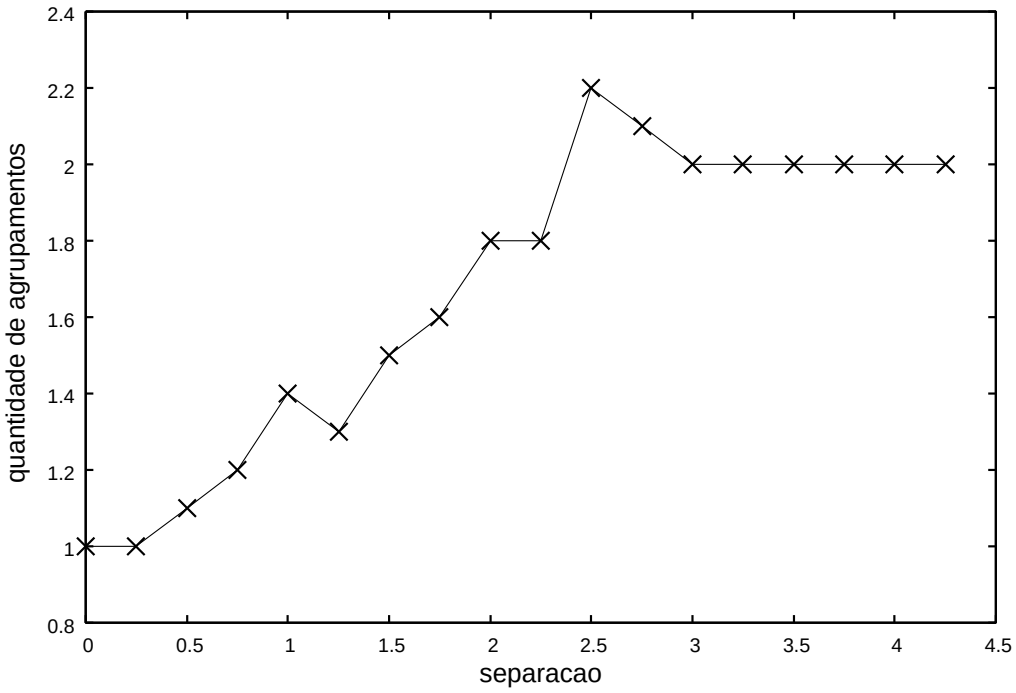


Figura 3.11: Avaliação da capacidade do algoritmo ABC para reconhecer agrupamentos independentes. Todos os bancos de dados usados foram gerados a partir de duas Gaussianas bidimensionais e desvio padrão unitário; cada agrupamento é composto por 1000 pontos. A separação é medida tomando como referência o desvio padrão. Para cada valor de separação, a quantidade de agrupamentos detectados é o valor médio de 10 experimentos independentes.

3.4.5 Escalabilidade com respeito à quantidade de atributos

O suporte celular empregado pelo algoritmo ABC é criado considerando, em cada operação de divisão celular, todos os atributos disponíveis. Cada célula dividida gera 2^d células filhas, sendo d a quantidade de dimensões do banco de dados. Portanto, o tamanho da lista que implementa o suporte depende exponencialmente da quantidade de atributos. No entanto, uma vez construído o suporte, a implementação do algoritmo somente considera todas as células no primeiro e no último passo da evolução do autômato, sendo que nos passos intermediários são consideradas apenas as células atratoras. Mas, ainda que somente seja necessário atualizar as células atratoras, as vizinhanças destas comportam uma quantidade de células que também depende exponencialmente de d. Portanto, a complexidade do algoritmo com respeito à quantidade de dimensões é de $O(2^d)$.

3.4.6 Escalabilidade com respeito à quantidade de agrupamentos

A quantidade de agrupamentos a serem detectados afeta fortemente o desempenho de alguns algoritmos. A influência da quantidade de agrupamentos no desempenho do algoritmo ABC e do CLUTO é comparado na Figura 3.12. Todos os bancos de dados empre-

gados nos experimentos contêm 100000 registros; eles foram gerados usando Gaussianas bidimensionais com desvio padrão unitário. Os centroides das distribuições foram uniformemente distribuídos de maneira que a menor distância entre dois deles quaisquer seja maior que 5 vezes o desvio padrão. Os tempos representados no gráfico correspondem aos valores médios de 5 execuções dos algoritmos. O tempo consumido pela leitura dos dados desde o arquivo não foram incluídos.

De acordo com o gráfico da Figura 3.12, o algoritmo ABC apresenta pouca variação no seu desempenho quando a quantidade de agrupamentos aumenta. O aumento no tempo médio consumido se deve a dois motivos: primeiro, a fragmentação é maior nos núcleos dos agrupamentos que nas periferias; isto implica num maior consumo de tempo na construção do suporte quando a quantidade de núcleos aumenta. Por outro lado, cada núcleo contribui com um conjunto independente de células atratoras que devem ser consideradas nos passos finais do algoritmo.

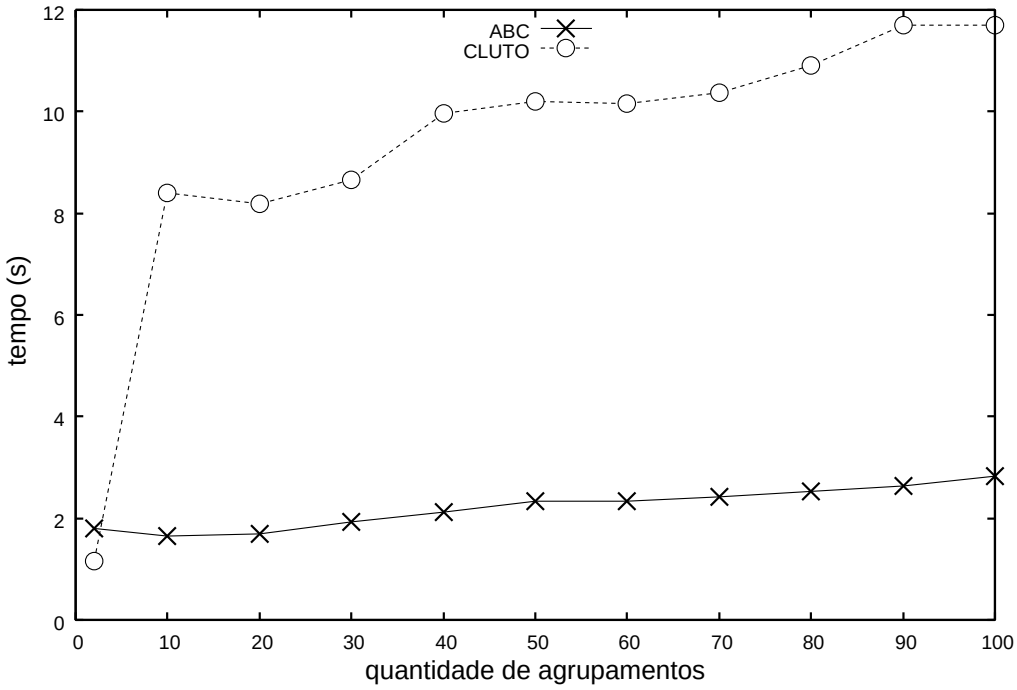


Figura 3.12: Comparação dos desempenhos do algoritmo ABC e do CLUTO com respeito ao tempo consumido na formação dos *clusters* em função da quantidade de agrupamentos no banco de dados. Os tempos correspondentes ao algoritmo ABC incluem à construção do suporte celular. Todos os bancos de dados estão compostos por Gaussianas bidimensionais. Todos os conjuntos de dados contam com 100000 registros.

3.4.7 Escalabilidade com respeito ao tamanho do banco de dados

Existem trabalhos (por exemplo, ver [1, 12]) que consideram separadamente o tempo consumido pelas etapas de pré-processamento e inicialização, do tempo empregado na detecção dos agrupamentos. No caso do algoritmo ABC, a etapa de construção do su-

porte celular acomoda essa informação na estrutura de dados sobre a qual a detecção dos agrupamentos é feita. Entretanto, consideramos que a construção do suporte está indissoluvelmente ligada ao funcionamento integral do algoritmo; por este motivo ela é sempre incluída no tempo de execução de ABC.

A construção do suporte celular é a fase que mais consome tempo durante a execução do algoritmo. O tempo consumido depende da distribuição de pontos do banco de dados, pois a etapa termina quando a diversidade geracional atinge o seu valor crítico[3]. A Figura 3.13 mostra a relação entre o tempo consumido para construir o suporte e o tamanho do conjunto de dados. Os tempos apresentados correspondem às médias de experimentos independentes. Os bancos de dados considerados foram gerados a partir da mistura de 5 Gaussianas bidimensionais, cada uma com diferente desvio padrão. O tempo medido não inclui a leitura dos dados desde o arquivo; ele compreende o lapso entre a criação da primeira célula e a formação do suporte com diversidade geracional crítica. O gráfico mostra que a complexidade algorítmica com respeito ao tamanho do conjunto de dados é $O(N)$, sendo N a quantidade de registros.

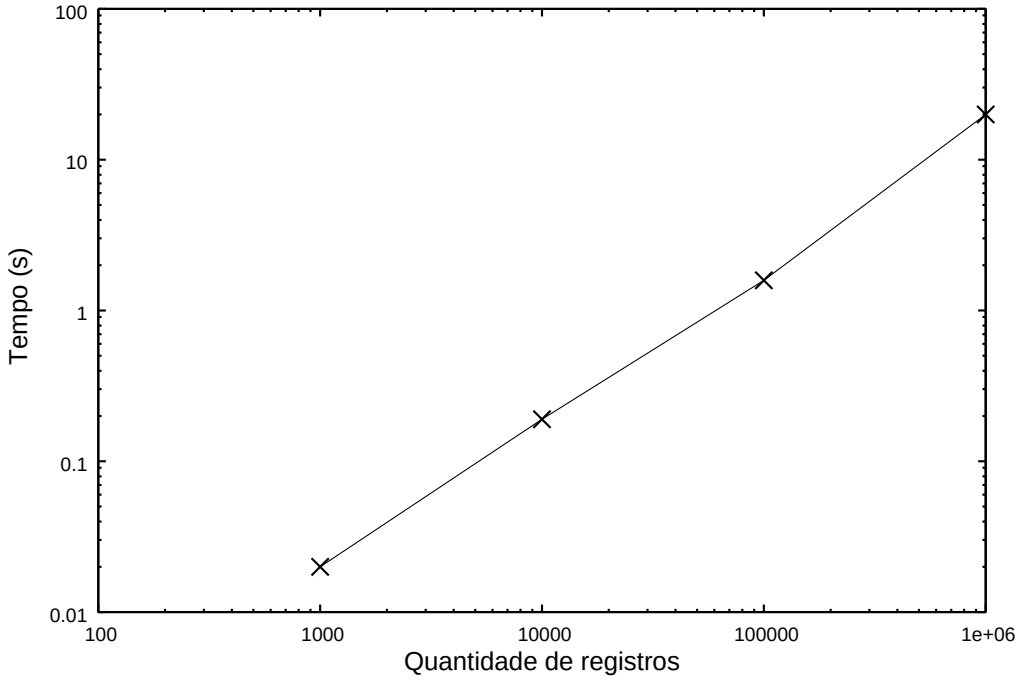


Figura 3.13: Relação entre o tempo consumido para construir o suporte e o tamanho do conjunto de dados. Os bancos de dados considerados são bidimensionais e compostos por uma mistura de Gaussianas com diferente desvio padrão.

A Figura 3.14 apresenta o tempo médio empregado, tanto pelo algoritmo ABC quanto pelo CLUTO, para a formação dos agrupamentos ao se considerar bancos de dados de diferentes tamanhos. Na medição desse tempo foram excluídas as operações de entrada e de saída de dados. Os bancos de dados utilizados nos experimentos estão compostos por

[3]O valor crítico da diversidade geracional Δg^* corresponde ao incremento relativo de densidade, ΔD_r^*.

uma mistura de 5 Gaussianas bidimensionais. O tempo consumido pelo algoritmo ABC é, para todos os conjuntos de dados considerados, aproximadamente a metade do tempo empregado pelo CLUTO. Para ambos os algoritmos, a complexidade algorítmica é $O(N)$, sendo N o tamanho do conjunto de dados.

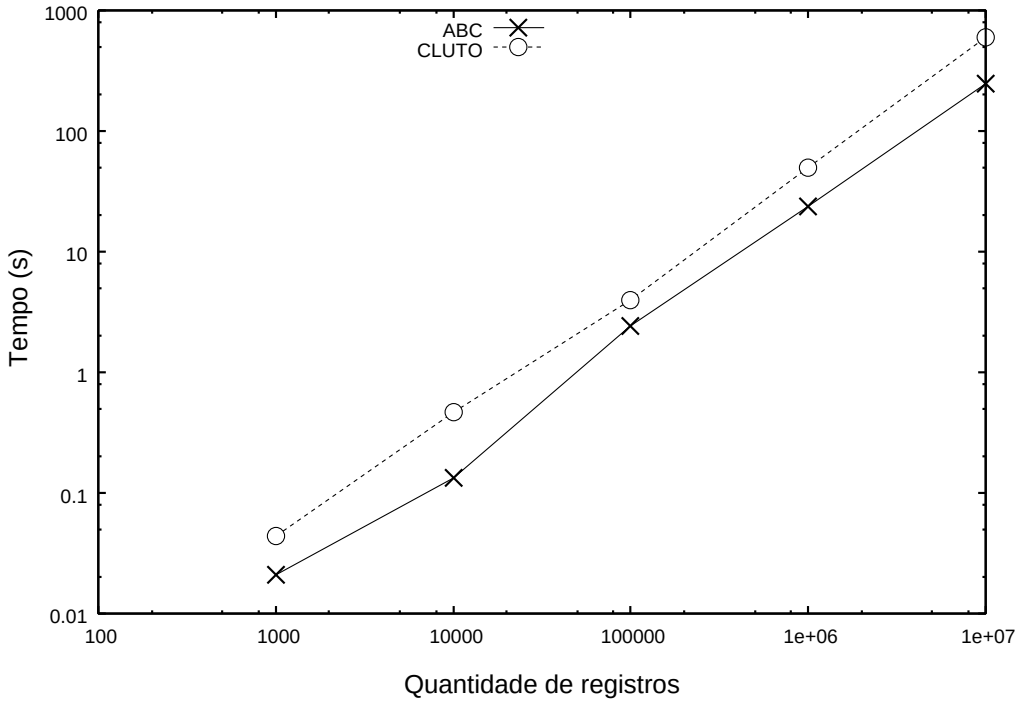


Figura 3.14: Comparação dos desempenhos do algoritmo ABC e do CLUTO com respeito à relação entre o tempo médio empregado para formar os agrupamentos e o tamanho do banco de dados. Os tempos correspondentes ao algoritmo ABC incluem a construção do suporte celular. Todos os bancos de dados estão compostos por 5 Gaussianas bidimensionais.

3.4.8 Desempenho em bancos de dados reais

Os desempenhos dos algoritmos CLUTO e ABC foram avaliados empregando dois bancos de dados disponíveis no repositório da Universidade da Califórnia em Irvine [2]. Os conjuntos de dados selecionados são: o gerado pelo telescópio para raios gama MAGIC[4], e o banco de dados STATLOG (Shuttle)[5]. A escolha se baseia nos seguintes critérios:

- eles possuem a série completa de valores para todos os atributos;
- todos os atributos são numéricos contínuos;
- eles contam com grande quantidade de registros (19020 no caso do MAGIC, e 43500 no conjunto de treinamento do STATLOG)

[4]http://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope
[5]http://archive.ics.uci.edu/ml/datasets/Statlog+%28Shuttle%29

Entretanto, devido à quantidade de atributos[6] (10 no MAGIC, e 9 no STATLOG), foi necessário pré-processar os dados antes de serem apresentados a ambos os algoritmos. O pré-processamento consistiu em considerar as séries de valores correspondentes às três componentes principais mais significativas. Desde que a quantidade de atributos não é uma limitação para o CLUTO, incluímos também os resultados desse algoritmo quando confrontado com os bancos de dados originais. Como já indicamos, o CLUTO requer que a quantidade de agrupamentos a serem detectados seja especificada de antemão, empregando-se para este propósito, a quantidade de classes em cada banco de dados. Entretanto, o restante dos parâmetros de CLUTO conservaram os seus valores por omissão.

Na avaliação foram considerados o tempo consumido na criação dos agrupamentos, e a taxa de erro da classificação resultante. Para calcular a percentagem de erros, a correspondência entre os rótulos das classes e os identificadores dos agrupamentos foi feita de modo que a cobertura correta de pontos seja máxima. As taxas de erro obtidas por ambos os algoritmos são grandes se comparadas com os valores observados no desempenho de algoritmos de aprendizado supervisionado.

A Tabela 3.2 contém os valores para o tempo e a taxa de erro obtidos na execução de ambos os algoritmos sobre o banco de dados MAGIC, e na Tabela 3.3 aparecem os correspondentes ao banco de dados STATLOG. A última coluna em ambas as tabelas corresponde ao desempenho do CLUTO sobre os respectivos bancos de dados originais (sem nenhum pré-processamento). No caso do banco STATLOG chama a atenção o baixo desempenho do CLUTO. O motivo para isto ter acontecido está na composição do banco de dados: uma das classes cobre 80% dos exemplos. O CLUTO usa a quantidade de agrupamentos fornecida como ponto de partida para a criação dos agrupamentos; isto tem por conseqüência uma fragmentação desnecessária da classe principal que deriva num incremento da taxa de erro. Já no caso do ABC, a quantidade de agrupamentos é resultado da execução do algoritmo; entretanto, o ABC somente detectou um agrupamento e por este motivo sua taxa de erro foi menor.

Todos os experimentos foram feitos usando um computador com processador Intel Core Duo 6300, com 3 GB de memória RAM, na plataforma Linux 32 *bits*. O algoritmo ABC foi implementado usando a linguagem de programação C++.

Tabela 3.2: Comparação dos desempenhos de ABC e de CLUTO, sobre o banco de dados MAGIC. A última coluna corresponde ao resultado da aplicação de CLUTO ao banco de dados original.

	ABC	CLUTO	CLUTO sem PCA
Erro (%)	35.72	44.01	32.30
Tempo médio (s)	0.38	0.41	0.75

[6]O atributo classe não foi considerado.

Tabela 3.3: Comparação dos desempenhos de ABC e de CLUTO, sobre o banco de dados STATLOG. A última coluna corresponde ao resultado da aplicação de CLUTO ao banco de dados original.

	ABC	CLUTO	CLUTO sem PCA
Erro (%)	21.6	67.1	66.1
Tempo médio (s)	1.05	2.95	3.39

3.4.9 Conclusões

Em síntese, o algoritmo ABC apresenta as seguintes caraterísticas operacionais:

- ele não possui parâmetros a serem sintonizados em cada aplicação concreta; por este motivo, ele é ideal para explorações iniciais em problemas de mineração de dados.

- as fronteiras dos agrupamentos têm limitada a sua resolução pela forma do suporte celular; entretanto, uma vez conhecida a classificação de pontos próximos aos núcleos dos agrupamentos, é possível empregar eficientes algoritmos de aprendizado supervisionado para definir as fronteiras inter-*cluster* de maneira ótima.

- a complexidade algorítmica com respeito à quantidade de atributos é $O(2^d)$, sendo d a quantidade de dimensões. No caso de aplicações com poucos ($\leqq 5$) atributos envolvidos, os dados não requerem qualquer tipo de pré-processamento. Para bancos de dados com grande quantidade de atributos é necessário incluir uma etapa prévia de seleção de atributos.

- a construção do suporte celular e o uso do gradiente de densidade como critério de agregação implicam numa baixa sensibilidade ao ruído, e na independência da ordem em que são considerados os registros.

- a complexidade algorítmica conjunta, que inclui a criação do suporte celular e a operação do autômato celular, com respeito ao tamanho do banco de dados é de $O(N)$, sendo N a quantidade de registros. Isto permite uma boa escalabilidade com respeito à quantidade de registros.

- o uso do autômato celular implica num processamento local da informação; por este motivo o desempenho do ABC é independente da quantidade de agrupamentos a serem detectados. Isto não é valido para algoritmos como o CLUTO.

- o processamento local da informação habilita naturalmente o algoritmo ABC para uma implementação paralela.

Os resultados obtidos com o uso do algoritmo ABC sugerem os próximos passos de aprimoramento:

1. desenvolver uma implementação paralela do algoritmo.
2. estender a construção do suporte celular para que possam ser considerados atributos categóricos.
3. considerar uma forma alternativa para a construção do suporte celular de modo que a complexidade algorítmica tenha dependência mais fraca da quantidade de atributos.
4. desenvolver uma implementação que integre uma etapa de otimização das fronteiras entre os agrupamentos. As máquinas de vetor suporte se apresentam como a primeira alternativa a considerar.

Capítulo 4

Interações Sociais

4.1 Introdução

O embasamento teórico desta parte da tese encontra-se nos trabalhos de dois autores: o sociólogo francês GABRIEL TARDE e o psicólogo canadense ALBERT BANDURA. Em oposição à concepção de sociedade de ÉMILE DURKHEIM, TARDE põe ênfase nos fenômenos socias como fruto das interações dos indivíduos que a compõem, colocando desta maneira em primeiro plano as influências recíprocas entre os elementos constituintes:

> "A expressão psicologia coletiva ou psicologia social é com freqüência entendida num sentido quimérico que é importante descartar desde o início. Ele consiste em conceber um espírito coletivo, uma consciência social, um nós, que existiria por fora ou por cima dos espíritos individuais. Não há qualquer necessidade, desde o nosso ponto de vista, desta concepção misteriosa para traçar entre a psicologia comum e a psicologia social – que nós estaríamos mais bem dispostos a chamar inter-espiritual – uma distinção bem nítida[1]." [73]

Para TARDE, o mecanismo dominante regulador do comportamento das massas é a *imitação* [74]. Mais recentemente, BANDURA sublinha a importância que tem para a educação das crianças a sua exposição repetida a cenas de violência, as quais seriam internalizadas e posteriormente atuadas devido ao mecanismo de imitação [5]. Destarte, influência

[1] "L'expression psychologie collective ou psychologie sociale est souvent comprise en un sens chimérique qu'il importe avant tout d'écarter. Il consiste à concevoir un esprit collectif, une conscience sociale, un nous, qui existerait en dehors ou au-dessus des esprits individuels. Nous n'avons nul besoin, à notre point de vue, de cette conception mystérieuse pour tracer entre la psychologie ordinaire et la psychologie sociale - que nous appellerions plus volontiers inter-spirituelle - une distinction très nette. Pendant que la première, en effet, s'attache aux rapports de l'esprit avec l'universalité des autres êtres extérieurs, la seconde étudie, ou doit étudier, les rapports mutuels des esprits, leurs influences unilatérales et réciproques - unilatérales d'abord, réciproques après. Il y a donc entre les deux la différence du genre à l'espèce ; mais l'espèce ici est d'une nature si singulière et si importante qu'elle veut être détachée du genre et traitée par des méthodes qui lui soient propres."[73]

e imitação estão, segundo o nosso ponto de vista, na base da formação de "estilos de comportamento" peculiares de cada sociedade. O comportamento criminoso é um caso particular de conduta que se espalha na sociedade devido à sua assimilação precoce, talvez pela concentração de maus exemplos (ver alelomimese em [37]), e certamente pela ausência de reforços exemplares dos comportamentos imprescindíveis para o convívio em sociedade. Atualmente o critério predominante para combater a criminalidade rampante se baseia na *lex mere penalis*: as regras que regulam a vida em sociedade são acatadas em função das penas que recaem sobre o infrator caso ele seja pego. Esta estratégia, baseada na vigilância e o castigo, é gulosa em recursos, tanto em quantidade de pessoas dedicadas ao monitoramento e à repressão, quanto em dinheiro público; os casos cada vez mais freqüentes de corrupção são apenas a manifestação do círculo vicioso, expresso coloquialmente na pergunta: "quem vai vigiar a esposa do vigia?" Certamente as instituções públicas que cuidam da criação e do respeito de leis são imprescindíveis, mas elas se baseiam no suposição de que o indivíduo conta com mecanismos internalizados para a vida em sociedade que fazem desnecessária a existência ubíqua de um agente repressor de infrações. A fiscalização e o policiamento são eficazes na medida em que se deparem com fatos delitivos esporádicos; eles podem restabelecer uma ordem pré-existente mas não podem ser o ponto de partida para a criação dessa ordem; eles podem compensar sintomas mas não são o alimento de uma sociedade saudável. A fonte da ordem social está nos próprios indivíduos que a compõem, na sua aptidão para a convivência social, e em última instância, naqueles mecanismos internalizados que levam o indivíduo a desconsiderar comportamentos que atentem contra o bem comum em benefício próprio. Estes mecanismos não estão inscritos no seu código genético mas eles são aprendidos desde cedo, no meio onde ele se desenvolve. Assim, seguindo as idéias de TARDE e BANDURA, acreditamos que a influência, como moduladora das relações socias, e a imitação, como mecanismo fundamental no processo de internalização de regras de comportamento, estão na base dos fenômenos socias emergentes. Por este motivo, a abordagem que adotamos para o estudo das relações sociais se baseia em indivíduos e suas interações, sendo esta abordagem *bottom-up* a mais adequada para entender o modo que uma sociedade se auto-organiza.

O objetivo deste capítulo é apresentar o modelo de AC que foi desenvolvido para representar de modo parcimonioso a dinâmica das influências que uma população mantém. Algumas características incorporadas se mostraram *a posteriori* suficientemente gerais para que com poucas modificações o modelo possa ser aplicado ao caso da evolução da criminalidade.

Nos Fundamentos Teóricos assinalamos que ainda não existe uma construção teórica que permita relacionar diretamente a *regra local* de mudança de estado com a *configuração global* do autômato. Portanto a construção de um modelo implica num processo iterativo com umas etapas guiadas pelo "bom senso" e a "intuição", e outras marcadas

pela experimentação. Talvez num último estágio corresponda fazer um estudo analítico que relacione a dinâmica local com alguma magnitude global, por exemplo, a distribuição de probabilidade dos estados das células.

Na Seção 4.2 são introduzidos os termos que serão empregados no modelo junto com uma definição formal; também descreve-se o modelo fornecendo a correlação das características consideradas com o nosso objeto de estudo. Na Seção 4.3 são apresentados resultados de experimentos computacionais que permitem conferir a consistência com as hipóteses que orientaram o desenho do modelo. Finalmente, na Seção 4.4 o modelo é aplicado ao estudo da evolução da criminalidade.

4.2 Definições e notações

O nosso objeto de estudo inicial é uma população de indivíduos e, nesse contexto, cada célula representa uma pessoa, analogamente ao feito por NOWAK e LEWENSTEIN desde o seu trabalho [54]. O tamanho da população é constante pois não se consideram nem mortes nem nascimentos. Nesse sentido correspondem diferentes interpretações para cada um dos elementos que compõem um AC: $\mathscr{A} = (\mathbb{Z}^d, \mathscr{S}, \mathscr{N}, \delta)$.

4.2.1 Reticulado e vizinhança

O modelo que apresentaremos possui um reticulado formado por um toróide bidimensional. Destarte, o conjunto de células se corresponde com um mini-mundo fechado, não sendo necessário considerar efeitos de borda.

A vizinhança de c_i corresponde ao grupo de pessoas que mantém interações significativas com ela. Designamos com $\mathscr{N}_F(c_i)$ à vizinhança do tipo MOORE com $r = 2$ para a célula c_i; o conjunto de células que compõem $\mathscr{N}_F(c_i)$ representa aqueles relacionamentos estáveis do indivíduo. Por exemplo, nesse conjunto estariam incluídos o seu núcleo familiar e o seu ambiente de trabalho. A maior parte dos estímulos significativos provém dessa componente fixa da vizinhança. No entanto um indivíduo está influenciado por outras fontes de informação que não estão sendo levadas em consideração na composição fixa de $\mathscr{N}_F(c_i)$; por exemplo, uma reportagem num meio de comunicação ou uma palestra podem influir na posição do indivíduo com respeito a um tema dado. Estas fontes instáveis de relacionamento equivalem ao que foi chamado *weak ties* por MARK GRANOVETTER (ver [6]). Elas seriam as responsáveis por fenômenos como o conhecido *small world* de STANLEY MILGRAM [50]. A forma escolhida para modelar estas influências de fontes variáveis de informação é ampliar a vizinhança para incluir células fora de $\mathscr{N}_F(c_i)$. Estas células, que representam interações não-locais, serão chamadas de *casuais*. Em cada passo de simulação elas são escolhidas em forma aleatória entre a população que não faz parte de $\mathscr{N}_F(c_i)$. A quantidade de casuais que compõe a vizinhança de uma cé-

lula qualquer é a mesma para todas as células. Nas simulações essa quantidade foi fixada como uma fração do tamanho de $\mathscr{N}_F(c_i)$ (ver [56]). Denotamos por $\mathscr{N}_V(c_i,t)$ o conjunto de células *casuais* que se incorporam à vizinhança de c_i no instante t. Deste modo, no instante t, a vizinhança de uma célula c_i é representada por

$$\overline{\mathscr{N}(c_i,t)} = \mathscr{N}_F(c_i) + \mathscr{N}_V(c_i,t) \tag{4.1}$$

Dado que o tamanho da vizinhança é o mesmo para todas as células, estamos supondo implicitamente que todos os indivíduos se relacionam com a mesma quantidade de pessoas. Esta simplificação representa a média para uma sociedade dada, e ela não significa um empobrecimento do modelo pois, como veremos adiante, o relevante para o modelo é a intensidade da resultante desses relacionamentos. Por isso, ainda que o número de relacionamentos no modelo seja o mesmo para todos os indivíduos, as conseqüências que esses relacionamentos têm variam de indivíduo para indivíduo. O tamanho da vizinhança é um parâmetro representativo do nível de extroversão de uma sociedade.

4.2.2 O conjunto de estados $\mathscr{S}$

A interpretação do valor do estado de um indivíduo é diversa e depende do fenômeno social a ser considerado. Por exemplo em [54] o valor do estado se corresponde com uma opinião referente a uma mudança no regime político. O conjunto de estados possíveis é binário pois nessa modelagem somente são consideradas as posições de adesão ou oposição à mudança política. Segundo NOWAK a polarização é uma tendência natural no caso de decisões muito relevantes para o indivíduo. No entanto o modelo que propomos tem a capacidade de capturar situações intermediárias que aceitem naturalmente uma diversidade maior de posições individuais. No modelo, o conjunto $\mathscr{S}$ é o conjunto $\mathbb{Z}$ dos inteiros. Pelas caraterísticas da regra local de mudança de estado que apresentaremos mais adiante, os estados das células ficaram restritos a um subconjunto finito. Na Seção 4.4 é apresentada uma aplicação do modelo; nela o estado representa a quantidade de crimes cometidos por um sujeito.

O estado de uma célula c_i no instante[2] $t \in \mathbb{N}$, é designado por $s(c_i,t)$.

4.2.3 A *pressão* sobre uma célula

Na Psicologia Social é usado o termo de *pressão* social -ou influência- para se referir ao modo como uma sociedade como um todo condiciona as condutas e atitudes que os seus membros podem adotar. Por exemplo em [56], esse conceito de pressão é empregado para modelar a coesão de um grupo. No presente modelo empregaremos um AC totalístico para

[2]Usaremos indistintamente o termo *instante* ou o termo *passo* para denotar uma iteração durante o curso de uma simulação.

representar o conceito de pressão. Assim, definimos *pressão* -ou *influência*- sobre uma célula c_i no instante t à expressão seguinte:

$$p(c_i,t) = \sum_{c_k \in \mathcal{N}_F(c_i)} s(c_k,t) + \sum_{c_k \in \mathcal{N}_V(c_i,t)} s(c_k,t) \tag{4.2}$$

Segundo a Equação 4.2, os estados de todas as células que compõem a vizinhança têm uma influência sobre c_i. Em alguns modelos vistos em outros trabalhos, a pressão leva em consideração um conceito de distância entre a célula alvo e suas vizinhas. A distância busca refletir o fato de que, para alguns fenômenos socias, as pessoas mais chegadas de um indivíduo teriam uma influência maior sobre ele que outras com as quais os vínculos são menos estreitos. Nesse caso a distância não representa uma distância física mas uma métrica sobre um reticulado de relacionamentos. KLÜVER *et al.* [40] interpretam a distância como o número de *hops*[3] numa rede de relacionamentos (*Social Network*).

Uma alternativa para modular a influência dos vizinhos é a empregada por NOWAK *et al.* (por exemplo ver [54]): para modelar o fato de que numa sociedade há pessoas mais influentes do que a média na formação de opinião, cada célula possui um atributo chamado *peso*. Esse peso pondera a contribuição que o estado de uma vizinha faz à pressão sobre a célula alvo.

No caso do modelo aqui apresentado, o próprio valor do estado é um modulador da influência: estados com valores absolutos grandes correspondem a indivíduos adotando posições radicais; estes indivíduos extremistas são mais propensos a querer influenciar nas posições de outros indivíduos mais centrados ou indiferentes.

4.2.4 Os limiares

Cada célula possui um par de limiares l (inferior) e u (superior) tais que $l, u \in \mathbb{R}$ e $l < u$. No início das simulações cujos resultados apresentamos, cada célula tem assinalados valores aleatórios[4] para u e l.

A função dos limiares é modelar a *anergia*[5] que um indivíduo mantém até que a pressão sobre ele atinge um valor que promove uma mudança nele. Dito de outro modo, nem todo estímulo provoca uma reação numa pessoa; o limiar é um recurso simples para separar aqueles estímulos inócuos daqueles que trazem implicações significativas na avaliação que a pessoa faz.

O fato de incluir dois limiares persegue o objetivo de criar um modelo geral o sufi-

[3] Se uma rede é representada por um grafo, um hop corresponde a uma aresta nesse grafo.

[4] Esse valores aleatórios são escolhidos seguindo uma distribuição normal sobre o intervalo de valores possíveis para a pressão. Esse intervalo é determinado pelo tamanho da vizinhança e pelos valores de s_min e de s_max.

[5] Termo usado como sinônimo de inatividade, passividade ou indiferença.

ciente para levar em conta que o estado pode evoluir em direções opostas em períodos diferentes para um mesmo indivíduo. Por exemplo, no trabalho de NOWAK citado anteriormente o estado é binário pois ele modela a adesão ou não a uma mudança política. Nesse caso um limiar é suficiente. No caso considerado na Seção 4.3 o estado não é binário, mas o sentido de variação do estado está fixado pelo fenômeno modelado, e portanto também consideraremos um limiar só. Ainda assim, aplicações futuras do modelo proposto considerarão fenômenos nos quais faz sentido manter dois limiares.

4.2.5 A diferença de estado Δs

Em cada passo t da simulação é calculada para cada célula c_i a *diferença de estado* $\Delta s(c_i,t) \in \{-1,0,1\}$ baseados nos valores da pressão e dos limiares:

$$\Delta s(c_i,t) = \begin{cases} 1 & u \leq p(c_i,t) \\ 0 & l < p(c_i,t) < u \\ -1 & p(c_i,t) \leq l \end{cases} \tag{4.3}$$

A diferença de estado é uma característica original do nosso modelo. Ela pretende capturar a característica dos sistemas vivos de regular suas mudanças seguindo gradientes. Por isso, do nosso ponto de vista, é biologicamente mais pertinente que o modelo preveja o sentido que uma mudança terá do que especificar *a priori* o valor do próximo estado. Esta abordagem se reconcilia com o tratamento clássico de fazer modelos baseados em equações diferenciais. A configuração global alcançada pelo AC traz pouca informação se pretende-se mapear os valores obtidos com magnitudes observadas no mundo real pois pequenas pertubações levam a estados globais muito diferentes. No entanto, concordando com a visão dos trabalhos de NOWAK *et al.*, é mais relevante poder estimar com uma precisão significativa o sentido em que mudarão as magnitudes reais baseados na dinâmica observada numa simulação do AC. Destarte, o resultado das simulações ganham em pertinência do ponto de vista prático como ferramenta de predição; também elas permitem confirmar se as hipóteses sobre a dinâmica local, ainda que sejam simplificações, levam em conta o que há de essencial na base de um fenômeno macroscópico complexo.

4.2.6 A regra local δ

4.2.6.1 Regra local determinística

Seja $s'(c_i,t+1) = s(c_i,t) + \Delta s(c_i,t)$. Para determinar o próximo estado $s(c_i,t+1)$ é empregada a seguinte regra local de atualização, $\delta : \mathscr{S} \longmapsto \mathscr{S}$:

$$s(c_i,t+1) = \begin{cases} \text{s\_min} & s^{'}(c_i,t+1) \leq \text{s\_min} \\ s^{'}(c_i,t+1) & \text{s\_min} < s^{'}(c_i,t+1) < \text{s\_max} \\ \text{s\_max} & \text{s\_max} \leq s^{'}(c_i,t+1) \end{cases} \tag{4.4}$$

onde s_min e s_max são dois parâmetros globais que definem o intervalo de valores permitidos para o estado de uma célula. A regra local apresentada acima somente completa o conceito de estarmos perante um modelo de AC baseado em diferenças. Entretanto, a regra local determinística, ainda que simples, não deixa espaço para outras contingências que podem influir de modo que a decisão seja adiada. Com o objetivo de levar em consideração esses fatores imponderáveis que levam a uma pessoa a reagir numa hora ou outra, consideramos uma regra local de mudança de estado *probabilística*.

4.2.6.2 Regra local probabilística

Uma regra local probabilística pode ser especificada levando em consideração uma das seguintes duas possibilidades:

1. A probabilidade de que a mudança aconteça é independente do estado atual do indivíduo. Neste caso o modelo deveria incorporar mais um parâmetro, seja ele P, que representa a probabilidade que uma célula mude o seu estado. Esse parâmetro poderia ser específico de cada célula -como acontece com os limiares- ou o mesmo para a população toda -como acontece com o tamanho da vizinhança. Esta última opção está na direção oposta da linha de raciocínio que nos levou a considerar a mudança probabilística pois, no caso de ser um parâmetro global, ele estaria caracterizando a população no seu conjunto e não o indivíduo. Portanto, para manter a coerência, suponhamos que essa probabilidade seja da forma $P(c_i)$. No entanto, esta escolha implica que o indivíduo não mude ao longo do tempo, ou seja, a disposição para mudar de estado é independente da sua história. Esta restrição pode ser uma aproximação adequada para cenários onde a escala de tempo seja breve se comparada com a duração da vida de um indivíduo. Com a finalidade de ganhar generalidade escolhemos para o modelo a opção descrita no ponto seguinte.

2. A probabilidade de que a mudança seja efetivada leva em consideração o estado atual do indivíduo. Uma consideração adicional é necessária para decidir como o estado pode participar na determinação dessa probabilidade. Observa-se que numa sociedade os grupos radicais tendem a ser minoria enquanto que a maior parte da população prefere adotar posições mais conservadoras perante as mudanças. Levando em conta este fato, a maneira mais evidente para relacionar o valor da probabilidade de mudança com o valor do estado é que um seja o inverso do outro. Destarte, valores maiores do estado, que representam posições mais radicais ou

extremas, têm uma probabilidade menor de serem adotadas. Por simplicidade, a primeira forma de calcular a probabilidade é fazê-la inversamente proporcional ao valor absoluto do estado.

A expressão da regra local dada pela equação 4.4 é adaptada para levar em consideração o raciocínio anterior. A nova regra δ é a seguinte:

$$s(c_i,t+1) = \begin{cases} s'(c_i,t+1) & \text{com probabilidade } \frac{1}{|s(c_i,t)|}\text{, ou se } s(c_i,t)=0 \\ s(c_i,t) & \text{no caso contrário} \end{cases} \tag{4.5}$$

Note-se que os parâmetros s_min e s_max podem ser dispensados já que estados com valores absolutos grandes são pouco prováveis; o novo conjunto de estados possíveis é $\mathscr{S} = \mathbb{Z}$. Esta regra probabilística simples tem a desvantagem de apresentar uma singularidade para o estado zero. Por esse motivo consideramos uma regra δ que fosse diferenciável afim de evitar artefatos nas simulações. Finalmente, a regra local adotada utiliza a função exponencial para calcular o valor da probabilidade. A Equação 4.5 é adaptada para levar em conta a consideração anterior, resultando na seguinte expressão:

$$s(c_i,t+1) = \begin{cases} s'(c_i,t+1) & \text{com probabilidade } \exp(-|\,s(c_i,t)\,|) \\ s(c_i,t) & \text{no caso contrário} \end{cases} \tag{4.6}$$

Notemos que a função exponencial poderia ser substituída por uma outra função diferenciável em $\mathbb{Z}$; por exemplo, poderia se escolher uma função polinomial. Entretanto, devido à falta de uma justificativa para uma escolha diferente à da função exponencial, os resultados computacionais apresentados na Seção 4.3 se baseiam no uso da Equação 4.6.

4.2.7 Limiares dinâmicos

Os limiares pretendem capturar a *sensibilidade* que uma pessoa possui perante a influência que o meio exerce sobre ela. Destarte, valores da pressão pertencentes ao intervalo $]l,u[$ são indiferentes para o indivíduo. No entanto, na realidade sabemos que na história de uma pessoa pode haver mudanças nessa atitude de indiferença que fazem que um mesmo estímulo que outrora fora inócuo, no presente provoque uma reação. Ainda que supor que a sensibilidade de um indivíduo é a mesma o tempo todo possa ser válido para alguns cenários – por exemplo, se a duração da simulação representa um período curto na vida de um indivíduo – um modelo mais geral deve considerar a possibilidade de que os valores dos limiares mudem ao longo do tempo.

Consideremos a seguir uma forma como os limiares de uma célula c_i podem ser modificados ao longo do tempo. Os organismos vivos tendem a funcionar com os seus parâmetros variando dentro de uma faixa, fenômeno este conhecido como *homeostase*. Ele pode ser visto como um equilíbrio dinâmico no qual pequenas perturbações são toleradas

para não ter a necessidade de considerar uma reação. Para isso, os seres vivos possuem dois mecanismos contrapostos para adaptar-se aos estímulos provenientes do meio.

O primeiro deles é chamado *acomodação* (α). Graças a ele um organismo deixa num segundo plano todos os estímulos provenintes do ambiente que comprometam a sua sobrevivência. Assim, uma pessoa se acostuma com o barulho da rua ou de um condicionador de ar e fica liberada para escutar aquilo que seja do seu interesse num dado momento. No caso dos limiares, esse mecanismo de acomodação faz com que eles mudem no mesmo sentido que Δs; ou seja, toda vez que a pressão atinja um nível tal que justifique uma mudança de estado, a célula se acomoda para que pequenas flutuações no valor da pressão não estejam provocando mudanças desnecessárias. Por exemplo, se o valor da pressão apresenta pequenas variações que atingem o limiar u, em lugar de mudar de estado o tempo todo, a célula muda o valor de u de modo a ficar insensível às flutuações.

Em contraposição ao mecanismo de acomodação está a chamada *sensibilização* (σ). Se uma fonte de estímulo desaparece, um organismo vivo procura se liberar do esforço desnecessário de estar reagindo ou filtrando estímulos inexistentes. Suponhamos que durante um lapso a pressão sobre c_i não atinge o seu limiar l; a sensibilização faz com que o valor de l mude progressivamente, atendendo ao fato de que não há mais necessidade de filtrar variações no valor da pressão.

No modelo vamos supor que os dois mecanismos supra-citados estão representados por dois parâmetros globais, $\alpha \in]1,+\infty[$ e $\sigma \in]0,1]$. Pelo fato deles serem globais, eles estariam caracterizando a população em geral, ainda que exista a possibilidade de que os seus valores sejam característicos do indivíduo. No entanto, devido à falta de argumentos concretos a favor desta última possibilidade, priorizaremos a parcimônia do modelo. No modelo, para atualizar o valor dos limiares, superpomos os efeitos da sensibilização e da acomodação: a sensiblização contribui com um termo propocional ao valor atual dos limiares, enquanto a acomodação aporta um termo proporcional à pressão além do limiar; quando a pressão não ultrapassa um limiar, o mecanismo de acomodação não é acionado. As equações que descrevem as adaptações dos limiares ao longo da simulação são as seguintes:

$$l(c_i,t+1) = \begin{cases} \sigma l(c_i,t) + \alpha[p(c_i,t) - l(c_i,t)] & \Delta s(c_i,t) < 0 \\ \sigma l(c_i,t) & \Delta s(c_i,t) \geq 0 \end{cases} \tag{4.7}$$

$$u(c_i,t+1) = \begin{cases} \sigma u(c_i,t) + \alpha[p(c_i,t) - u(c_i,t)] & \Delta s(c_i,t) > 0 \\ \sigma u(c_i,t) & \Delta s(c_i,t) \leq 0 \end{cases} \tag{4.8}$$

O autômato celular desenvolvido para modelar interações sociais possui as caraterísticas resumidas a seguir:

- Ele é totalístico pois segundo a Equação 4.2 é considerada a soma dos estados das

células vizinhas.

- Ele é diferencial desde que a Equação 4.3 especifica a variação de estado que deve ser usada para calcular o estado seguinte.
- Ele é estocástico pois o próximo estado de uma célula é determinado seguindo a regra local probabilística especificada por 4.6.
- Ele é híbrido pois cada célula tem a sua regra local dependente do valor dos limiares, sendo que estes são diferentes de uma célula para outra; dito de outro modo, ainda que a estrutura da regra local seja a mesma para todas as célula, sua expressão concreta é diferente para cada uma delas.
- Ele é programável pois a regra local depende do valor dos limiares, sendo que estes mudam ao longo do tempo.

Todas as caraterísticas anteriores fazem do autômato celular proposto um modelo apropriado para representar o comportamento de um sistema complexo adaptativo.

4.2.8 O algoritmo de simulação

O algoritmo empregado para efetuar as simulações é o seguinte:

Algoritmo 6 Algoritmo de simulação do AC probabilístico, com limiares variáveis.

```
inicializar o estado e os limiares de todas as células
para cada passo faça:
    para cada célula faça:
        calcule a pressão sobre a célula
        calcule a diferença de estado
        calcule o próximo estado
        calcule o valor dos limiares
```

4.3 Experimentos computacionais

Nas seções anteriores foram descritas as caraterísticas de um autômato celular que tem por objetivo modelar as interações entre indivíduos e as suas conseqüências. As caraterísticas devem levar a configurações globais consistentes com a auto-organização observável numa sociedade qualquer. Uma série de experimentos computacionais foram feitos a fim de conferir que o comportamento global do AC se corresponde com as hipóteses assumidas no seu projeto. A seguir apresentamos uma série de figuras que mostram que o comportamento do AC é consistente com o esperado.

Todos os experimentos computacionais foram feitos com os seguinte valores para os parâmetros:

- O tamanho do reticulado bidimensional toroidal é de 50×50 células.
- Todas as células são inicializadas com o estado nulo.
- Os limiares para cada célula são inicializados com valores aleatórios segundo uma distribuição uniforme de valores. O intervalo de valores iniciais foi determinado experimentalmente de modo que eles sejam atingidos pelo valor da pressão em experimentos com duração de 500 passos. Ou seja, o intervalo é largo o suficiente para que a população apresente diversidade, mas não inclua valores tão grandes para que não afetem a dinâmica durante a duração do experimento.
- O tamanho de $\mathscr{N}_V(c_i,t)$ (conjunto das vizinhas casuais) é a quinta parte do tamanho de $\mathscr{N}_F(c_i,t)$. Fixado o raio de $\mathscr{N}_F(c_i,t)$ no valor 2, cinco vizinhas casuais são incorporadas à vizinhança de cada célula, sendo que elas são escolhidas aleatoriamente em cada passo de simulação. A escolha é feita segundo uma distribuição de probabilidade uniforme.

A avaliação do comportamento global do AC é feita usando os indicadores seguintes:

- Em cada passo de simulação um subconjunto da população inteira de células muda o seu estado. O tamanho desse subconjunto representa a *quantidade de mudanças por passo*. A variação deste indicador ao longo do tempo mostra se o sistema tende a uma configuração de equilíbrio dinâmico.
- Cada célula reage à pressão à qual é submetida de duas maneiras: mudando o seu estado e mudando os valores dos seu limiares. Para cada célula é guardada a evolução dos valores da pressão e de seus limiares.
- A configuração global do AC pode ser avaliada observando a distribuição de freqüência do valor do estado na população. Lembremos que inicialmente todas as células são inicializadas com o estado nulo; no entanto, os valores aleatórios dos limiares de cada célula induzem algumas células a mudar de estado. Destarte, o sistema se afasta da configuração inicial uniforme, onde todas as células têm o mesmo valor de estado.

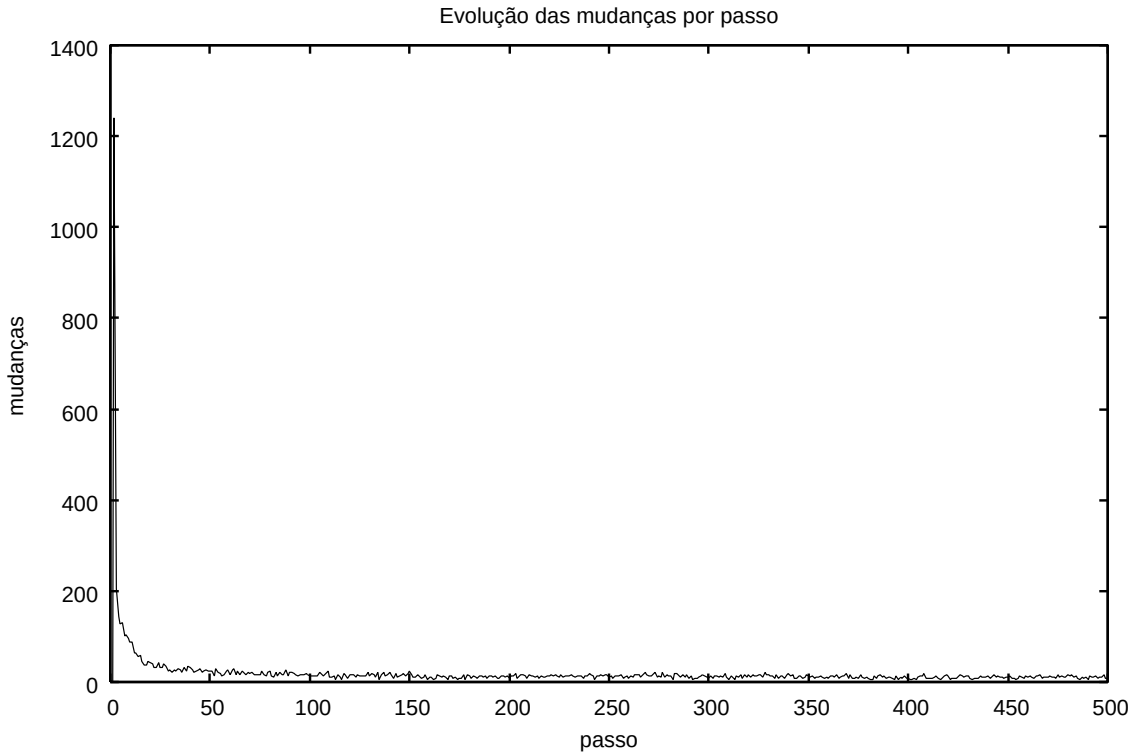


Figura 4.1: Evolução da quantidade de mudanças por passo

As figuras apresentadas a seguir foram escolhidas por serem representativas do observado em numerosas simulações independentes. Na Figura 4.1 está o gráfico da evolução da quantidade de mudanças por passo. Após um breve transitório, a quantidade de mudanças oscila em torno de um valor médio constante. Esse valor médio sugere que quando o sistema atinge o regime estacionário, somente um subconjuto reduzido de células muda o seu estado. Essas mudanças têm a sua origem na combinação dos efeitos da sensibilização e das interações casuais entre células. Este equilíbrio dinâmico é consistente com o comportamento observado numa sociedade, onde permanentemente há indivíduos que mudam suas opiniões ou escolhas. Esta configuração de equilíbrio dinâmico é atingida como conseqüência da auto-organização do sistema.

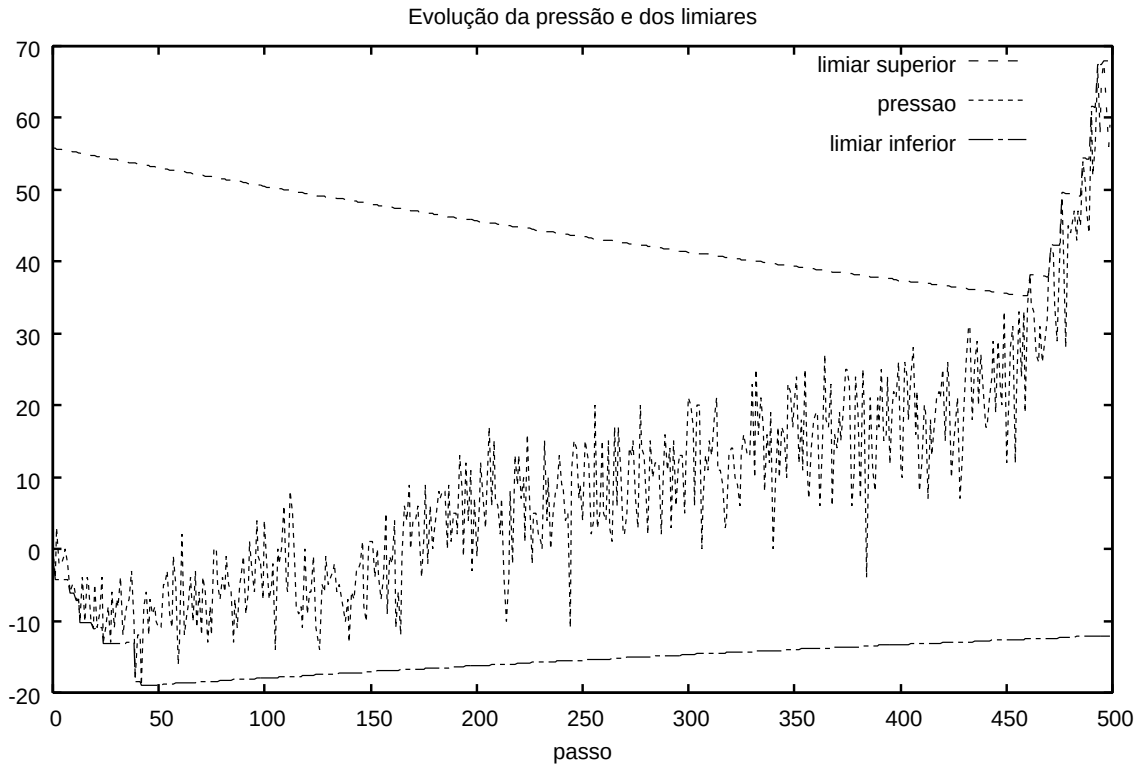
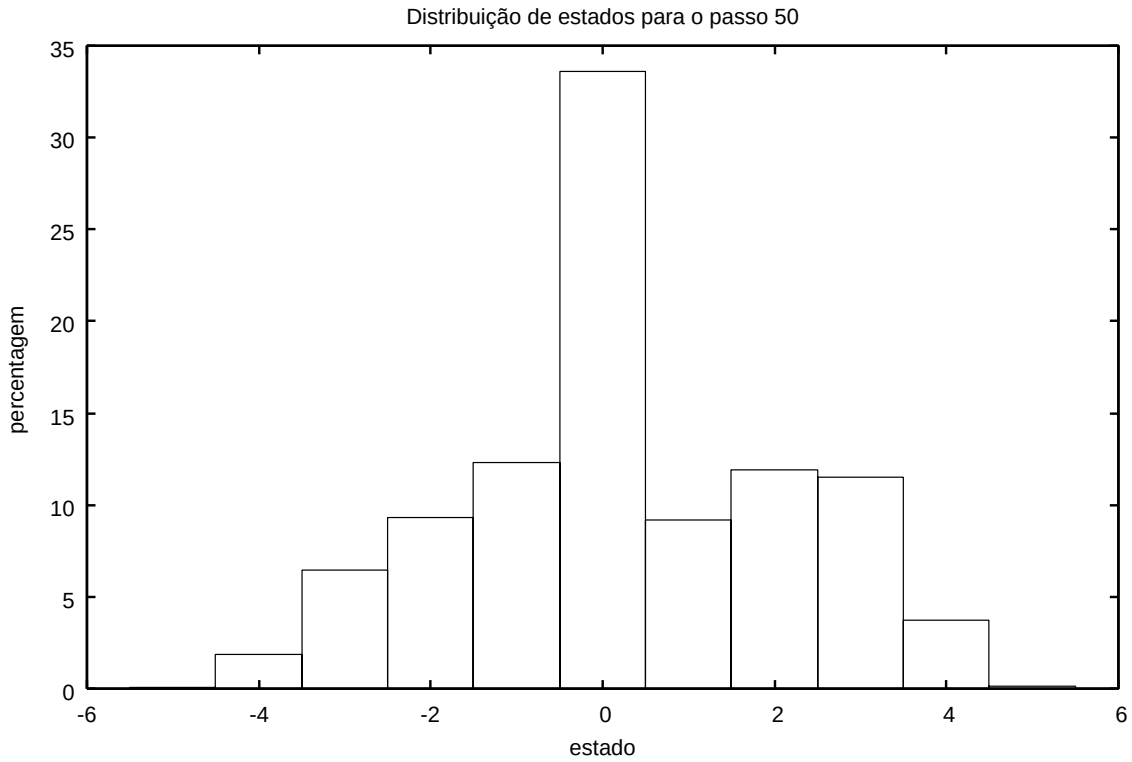


Figura 4.2: Exemplos de evolução da pressão.

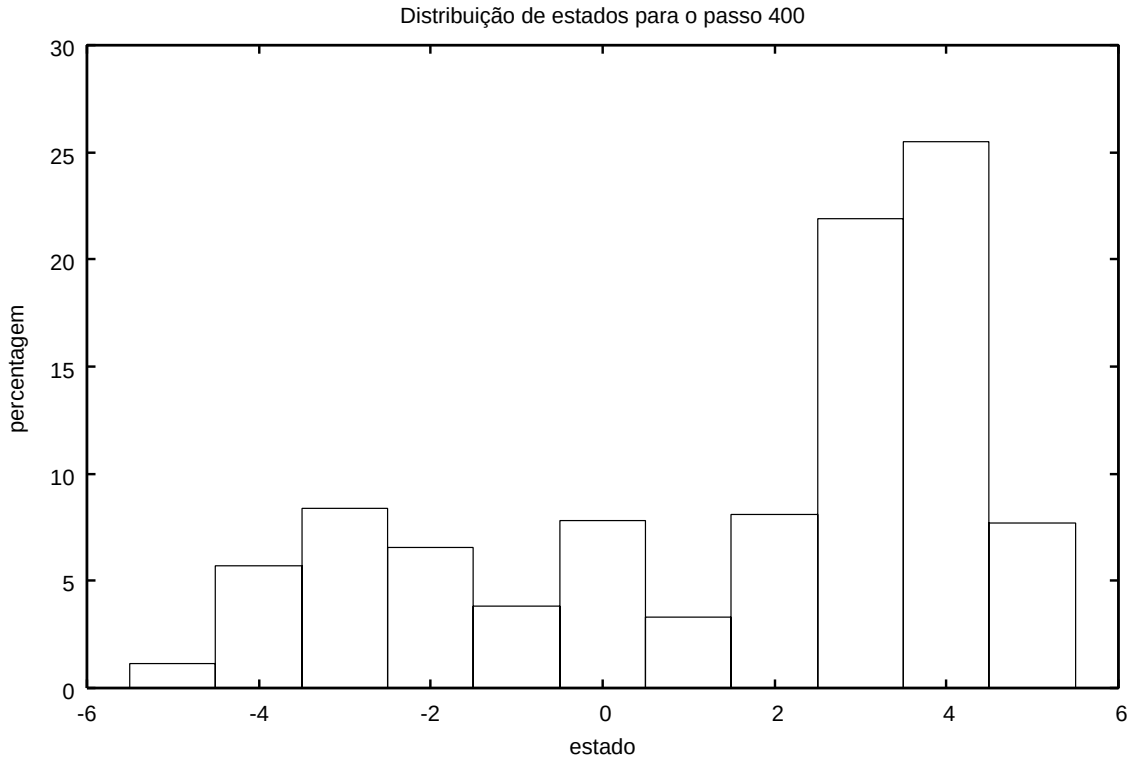
Na Figura 4.2 é apresentado o gráfico com a evolução do valor da pressão e dos limiares para uma célula qualquer. Ficam em evidência os mecanismos de sensibilização -o limiar superior no início da simulação- e de acomodação -no limiar inferior, no mesmo período. Na configuração de equilíbrio, a maior parte das células terão adaptados os seus limiares de modo a ficarem indiferentes à pressão das vizinhas; nessa situação cada célula gastaria o mínimo de energia[6] nas interações com o restante das células. Em biologia, esse estado de indiferença ou de ausência de reação é chamado de *anergia*. Os limiares têm um efeito amortecedor na dinâmica do sistema. Segundo JENSEN [35], a combinação de pequenas mudanças -esse é o intuito de considerar um AC diferencial- junto com limiares locais, por exemplo, a fricção estática- está na base de sistemas criticamente auto-organizados; no entanto, no presente modelo não foi incluido nenhum fator externo[7] que presione uniformemente a população toda e por esse motivo não emergem configurações críticas.

[6]Por simplicidade no modelo não é considerada nenhuma restrição do tipo energético. No entanto, é sensato supor que toda mudança de um ser vivo com o objetivo de se adaptar ao meio implica no consumo de algum recurso.

[7]Por exemplo, em alguns sistemas criticamente auto-organizados como pilhas de arroz, a força gravitacional atua uniformemente sobre todos os elementos do sistema.



(a)



(b)

Figura 4.3: Exemplos de distribuições de estado

As Figuras 4.3 a) e b) apresentam o histograma de freqüências dos valores dos estados para dois instantes de um experimento qualquer. O intervalo de valores é auto-limitado

pela escolha da regra local probabilística que torna improváveis os valores extremos. A distribuição de freqüências se uniformiza ao longo do tempo ainda que a tendência seja para uma distribuição bimodal. O motivo para isto está na expressão 4.2 para o cálculo da pressão: a configuração global no equilíbrio dinâmico tende a minimizar os valores da pressão sobre cada célula, e por isso os estados com valor positvo tendem a ser compensados com estado com valor negativo. Deste modo, o autômato celular totalístico captura naturalmente a tendência à polarização das opinões assinalada por NOWAK.

4.4 Aplicação do modelo

O presente trabalho pretende seguir as linhas conceituais contidas nos trabalhos de NOWAK ([45][53][54]) e de TURNER *et al.* [77]. De [54] resgatamos a ênfase posta na modelagem como mecanismo de compreensão qualitativa dos fenômenos sociais. O segundo é um exemplo de como as visões "micro" e "macro" convergem na produção de resultados.

O trabalho de COOK, ORMEROD e COOPER *"Scaling Behaviour in the Number of Criminal Acts Committed by Individuals"* (ref. [18]) é um estudo baseado em dois bancos de dados independentes sobre as atividades criminosas de dois grupos de indivíduos ao longo de períodos que vão de meses até mais de uma década. As características desses bancos de dados os fazem particularmente interessantes tanto do ponto de vista estatístico (pelo volume e a representatividade das amostras de indivíduos considerados) assim como do ponto de vista social pelo acompanhamento desses grupos de pessoas ao longo do tempo. A conclusão mais importante do trabalho se refere ao fato de que em ambos os bancos de dados observa-se uma relação do tipo *lei de potência* entre a quantidade de delitos cometido por um mesmo indivíduo e a quantidade de indivíduos que coincidem nessa quantidade de crimes.

Distribuições de freqüência do tipo lei de potência são achadas nos mais diversos tipos de fenômenos na natureza (ver por exemplo [3, 35]). Elas estão associadas a dinâmicas de sistemas complexos nos quais há algum tipo de interação cooperativa entre os elementos constitutivos. Freqüentemente essas distribuições são indicativas de organizações dentro da população de elementos que independem da escala ou tamanho total do sistema.

No nosso caso procuramos achar uma relação do tipo

$$\text{freqüência}(s) = \kappa s^{\beta} \tag{4.9}$$

onde s é o valor do estado de uma célula, e o expoente β é característico da dinâmica, e independe do tamanho do sistema considerado. Note-se que se a relação é representada com escala logarítmica em ambos eixos, ela corresponderá a uma reta.

4.4.1 Adaptações feitas no modelo

No modelo que apresentamos no Capítulo 3 estão potencialmente presentes as características de um sistema complexo com tendência a auto-organizar-se. Portanto, decidimos conferir se é possível fazer nele adaptações para aproximá-lo às características estudadas em [18].

A primeira adaptação é supor que o estado de uma célula c_i corresponde ao número de crimes cometidos pelo indivíduo representado pela célula. Como conseqüência desta escolha, o conjunto de estados possíveis fica determinado: $\mathscr{S} = \mathbb{N}$. Desde que a quantidade de crimes cometidos não pode diminuir, não faz sentido considerar o limiar inferior l, e deixamos o seu valor fixado em zero.

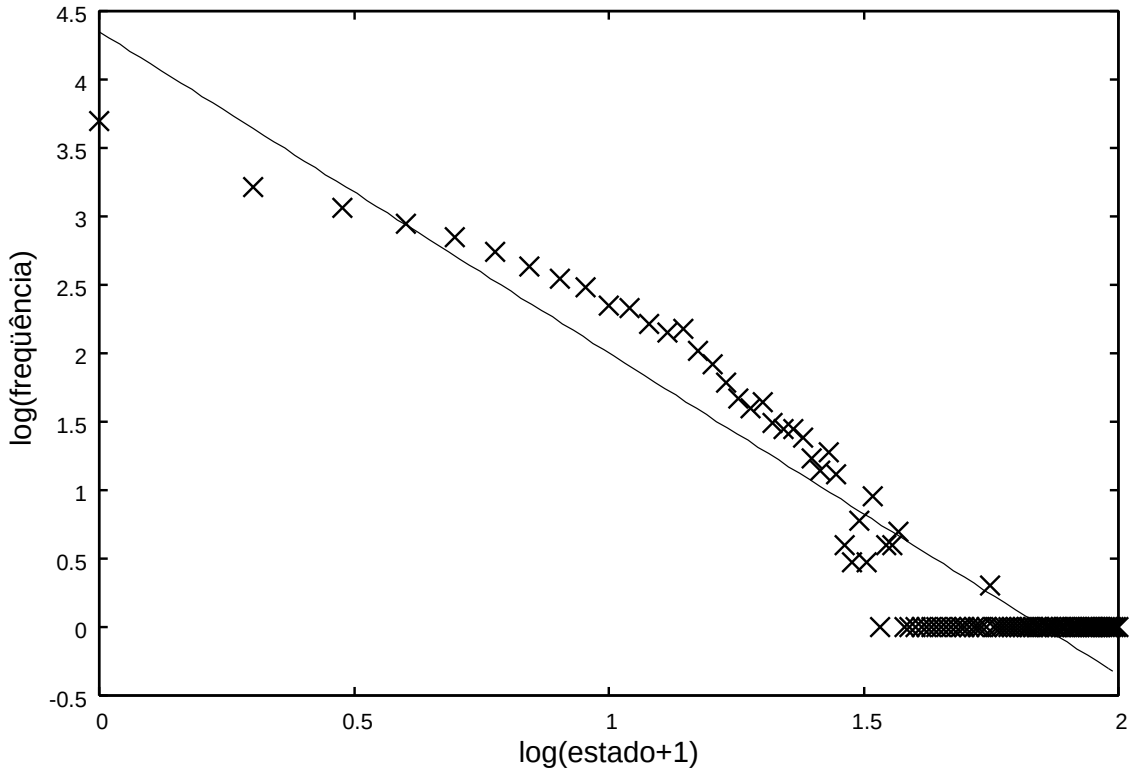


Figura 4.4: Gráfico log-log da distribuição de estados

Os valores para os parâmetros foram os seguintes:

- tamanho do reticulado toroidal: 200×200.
- raio da vizinhança: 1
- quantidade de *casuais:* 3
- sensibilização $\sigma = 1$
- acomodação $\alpha = 1.10$

- total de passos de simulação: 5000
- o valor inicial para o estado das células é 1 com uma probabilidade de 10%, ou zero no caso contrário.
- o valor inicial do limiar *u* de cada célula é assinalado segundo uma distribuição de probabilidade uniforme sobre o intervalo $[0, 100]$

4.4.2 Resultados preliminares

Na Figura 4.4 aparece o gráfico log-log[8] da freqüência de células versus o valor do estado[9]. No gráfico está também representada a reta que melhor aproxima a distribuição de pontos[10]. Nota-se que o espalhamento da nuvem de pontos para os grandes valores do estado é uma conseqüência do tamanho relativamente reduzido do reticulado escolhido.

Há uma forte semelhança entre o gráfico obtido dos resultados da simulação e os dados apresentados em [18]. O coeficiente do ajuste linear no nosso caso é de $\beta = -2.34$ com um erro padrão de 0.077 enquanto COOK *et al*. obtêm valores de -2.28 e 0.071. A proximidade dos valores pode ser apenas resultado de uma coincidência, pois não houve nenhum cuidado especial na escolha dos parâmetros de simulação para nos aproximar aos resultados apresentados em [18]. O propósito das simulações feitas é explorar a possibilidade do nosso modelo capturar uma dinâmica como a descoberta em [18]. No estudo preliminar somente foi pesquisada a pertinência das características que o nosso modelo incorpora. Nesse sentido pode ser afirmado que tanto a decisão probabilística quanto o mecanismo de sensibilização são as duas características necessárias ao modelo para que distribuições do tipo lei de potência possam ser achadas. Seguindo a linha de raciocínio de NOWAK, é preciso aprofundar nas correlações que tais características mantêm com a realidade para obter uma compreensão melhor do fenômeno modelado.

Curiosamente, a partir dos resultados preliminares das simulações, podem ser feitas considerações análogas às achadas em [18] com respeito da tendência bimodal da distribuição; nesse sentido, se desconsideramos o grupo de células com estado 0 (que se corresponde a indivíduos que nunca cometeram um delito), o ajuste a uma lei de potências da distribuição de freqüência dos estados melhora muito. COOK *et al*. reconhecem nesta observação a existência de um passo crucial na dinâmica do comportamento delituoso: desde que o indivíduo cometa o primeiro crime, a quantidade de crimes que ele poderá vir a cometer parece não ter limite de escala.

[8] Foi empregado o logaritmo de base decimal.

[9] Estritamente, nas abcisas do gráfico aparece o valor do estado incrementado de um. O motivo desta translação da origem é possibilitar a representação do logaritmo do estado para aquelas células com estado 0. Este recurso também é empregado no referido artigo de COOK *et al*., onde chamam de *rank* de um indivíduo à quantidade de crimes por ele cometidos acrescentada de um. Por simplicidade decidimos não seguir a terminologia de COOK *et al*.

[10] Para a estimação do coeficiente da reta foi empregado o algoritmo fornecido pelo programa Gnuplot, baseado no algoritmo de MARQUARDT-LEVENBERG para *Nonlinear Least-Squares* (NLLS).

Capítulo 5

Conclusão

Os sistemas complexos carecem de uma definição e tratamento formal dentro da teoria dos sistemas dinâmicos. Para o seu estudo empregam-se modelos computacionais que colocam o foco nos elementos componentes e na maneira como estes interagem. Os autômatos celulares são uma ferramenta conceitual particularmente adequada para o estudo de sistemas complexos pois eles baseiam a sua operação na aplicação de regras simples em pequenos subconjuntos (vizinhanças) dos elementos (células) que os integram. Os fenômenos não-lineares observados nos sistemas complexos podem ser modelados concentrando a sua origem na estrutura da regra local, sem afetar as características dos elementos. Alternativamente, como fazemos no presente trabalho, a não-linearidade pode ser modelada como resultado de regras locais simples aplicadas a elementos cujas caraterísticas mudam ao longo do tempo como resultado da história de interações. Assim, estas caraterísticas podem ser interpretadas como variáveis de integração de um potencial, ou poder ser vistas como uma espécie de memória distribuida em cada elemento. Devido à participação destas caraterísticas na dinâmica das interações, designa-se-as com o termo genérico influência.

Para os sistemas dinâmicos descritos a partir de equações diferenciais existem critérios analíticos que permitem caraterizar a forma das trajetórias no espaço de estados: é possível predizer a existência de atratores, assim como prever a estabilidade das trajetórias na proximidade destes atratores. Em contrapartida, no estudo dos autômatos celulares, apresenta particular interesse a relação entre a definição da regra local e o conjunto de configurações globais atingíveis pelo autômato. Com exceção de alguns casos triviais, não é possível determinar analiticamente se uma configuração global pode ser alcançada por um autômato, dados os estados iniciais das células e a regra local. A única alternativa para conhecer as configurações globais exatas possíveis de um autômato é simular a sua evolução, passo a passo. Entretanto, em alguns casos, após muitas simulações, é possível ganhar intuição sobre as caraterísticas das configurações globais atingíveis por um autômato dado.

O projeto de um autômato celular cujas configurações globais venham a ser de uti-

lidade para um propósito específico implica num longo processo de experimentação a consequência do qual, aos poucos, “emerge” uma compreensão da relação entre os fenômenos locais (sobre os quais detemos o poder de especificar) e o comportamento do sistema como um todo. O desafio do projeto do autômato celular determinístico, de quem deriva o algoritmo de *clusterização*, consistiu em amadurecer essa compreensão, que uma vez presente, permite reconhecer os princípios gerais que regem a sua dinâmica global.

A aplicação prática mais comum dos autômatos celulares é a modelagem de um sistema complexo concreto com o propósito de identificar as principais variáveis, assim como as suas inter-relações, que estão na base da emergência da dinâmica global observada no sistema. Os fenômenos que aparecem num sistema complexo não estão determinados unicamente pela ação de uma força ou de uma restrição alheia ao sistema. Esses fenômenos têm a sua origem distribuída nas interações entre os elementos constituintes do sistema; por este motivo, o foco do estudo de um sistema complexo está na dinâmica destas interações. Entretanto, a relação entre as caraterísticas destas interações e os fenômenos globais, geralmente está longe de ser evidente. Nasce assim a necessidade de explorar, usando a simulação de modelos, como uma dinâmica local concreta repercute no comportamento global. O autômato celular estocástico criado com a finalidade de estudar fenômenos sociais procura considerar aquelas caraterísticas locais que consideramos primárias, para que o modelo possa exibir comportamentos globais análogos aos observados na realidade.

Os dois autômatos celulares apresentados compartilham o intuito de mostrar que a generalização feita, introduzindo o conceito de influência, permite modelar um sistema que se auto-organiza, levando em consideração que os elementos componentes se comportam de acordo com a sua história de interações. O papel da influência pode ser visto como o dual do peso das conexões sinápticas das redes neuronais artificiais. A influência como propriedade dinâmica dos elementos incrementa a capacidade de memorizar do sistema, que nos autômatos celulares clássicos se reduz à distribuição dos valores dos estados das células.

No restante do capítulo fazemos a síntese dos méritos principais dos autômatos desenvolvidos e indicamos possíveis direções para trabalhos futuros.

Autômato celular generalizado para a detecção de agrupamentos

O autômato apresentado no Capítulo 3 é o substrato usado para o desenvolvimento do algoritmo de *clusterização* para grandes bancos de dados. As suas vantagens são:

- os dados podem ser processados sem a necessidade de fazer algum tipo de normalização, mudança de sistema de coordenadas etc.
- não é necessário especificar o valor de parâmetros, o que torna o algoritmo adequado para uma primeira exploração em mineração de dados.

- como corolário do item anterior, a quantidade de agrupamentos decorre do processo de detecção.
- a construção do suporte celular dispensa do uso de uma métrica para avaliar a distância entre pontos. A proximidade é avaliada através das vizinhanças definidas sobre o suporte. As operações envolvidas na construção do suporte implicam em simples subtrações de magnitudes: o custo computacional por este conceito é mínimo.
- o algoritmo não faz uso de nenhum modelo probabilístico a ser calibrado com a distribuição dos pontos. Os modelos probabilísticos geralmente implicam em pesados cálculos para a estimação de parâmetros.
- não são empregadas transformações computacionalmente custosas como por exemplo a convolução com um *kernel*, ou a aplicação de *wavelets.*
- as células – e não os pontos – são as entidades processadas pelo algoritmo; desde que cada célula pode conter várias centenas de pontos cada, o algoritmo apresenta escalabilidade com respeito à quantidade de registros a processar. Técnicas de amostragem, frequentemente usadas por outros algoritmos para lidar com grandes bancos de dados, são opcionais mas não necessárias.
- os agrupamentos são detectados independentemente da sua forma, incluindo o caso de *clusters* que não são linearmente separáveis. A presença de ruído na forma de *outliers* tem um impacto mínimo sobre a qualidade da classificação.
- o uso do autômato celular generalizado implica num tratamento localizado da informação; por este motivo, o algoritmo – assim como a própria construção do suporte celular – apresenta uma disposição natural para uma implementação paralela: o custo computacional pode ser distribuído entre vários processadores, aumentando assim sua eficiência.

Entretanto, o algoritmo proposto possui duas grandes desvantagens:

1. a definição pobre das fronteiras dos agrupamentos.
2. a complexidade algorítmica com respeito da quantidade de atributos é de ordem exponencial.

A primeira desvantagem é consequência do mapeamento da nuvem de pontos que representa o banco de dados na estrutura do suporte celular. As periferias dos agrupamentos são regiões de baixa densidade e portanto, estão associadas a células de maior volume. Com isto, a definição da fronteira entre dois agrupamentos tem a sua resolução comprometida. No entanto, a classificação dos pontos da periferia dos agrupamentos representa

um desafio para todos os algoritmos. A detecção de agrupamentos pertence à categoria dos problemas de aprendizado não-supervisionado, e portanto, na ausência de etiquetas que orientem à decisão de pertença, a classificação final de um ponto fronteiriço comporta sempre um quota de arbitrariedade. No entanto, um dos objetivos do *clustering* é gerar partições nas quais os agrupamentos maximizam a separação entre eles. Por este motivo, é possível melhorar a definição das fronteiras entre os agrupamentos detectados pelo nosso algoritmo, incluindo uma etapa adicional que transforme o problema da definição num caso de aprendizado supervisionado. Para isso, podemos considerar que os pontos mais próximos dos núcleos dos agrupamentos, cuja classificação é certa, representam exemplos de classes; nesse contexto, o passo seguinte é treinar classificadores tais que maximizem a separação entre as classes, ou equivalentemente, os agrupamentos. Numa etapa seguinte, esses classificadores são empregados para decidir a pertinência dos pontos fronteiriços e por conseguinte, delineando as fronteiras com a maior acurácia. As *máquinas de vetores de suporte* [67] são as candidatas ideais para gerar esses classificadores. Notemos entretanto que não é necessário empregar todos os pontos pertencentes ao núcleo de um agrupamento: apenas aqueles pontos mais externos do núcleo são necessários pois entre eles se encontram os vetores de suporte. Destarte, um trabalho futuro a considerar é estudar a forma de implementar esta etapa de definição acurada das fronteiras entre os agrupamentos empregando máquinas de vetor de suporte.

A segunda desvantagem do nosso algoritmo pode ser contornada considerando duas estratégias alternativas. A primeira estratégia consiste em incluir uma etapa prévia à construção do suporte celular, na qual é feita uma seleção dos atributos mais relevantes a serem considerados para fazer a *clusterização*. Esta estratégia é empregada por outros algoritmos de detecção de agrupamentos e a técnica é conhecida como *feature selection*.

A outra estratégia a ser explorada envolve uma mudança na maneira como o suporte celular é construído, adaptando se for necessário, os critérios fornecidos para determinar o grau de fragmentação (diversidade geracional) e tamanho máximo da vizinhança. A bipartição de todas as dimensões do banco de dados pode ser desnecessária no caso daqueles atributos com uma distribuição uniforme de valores sobre o intervalo a dividir. Deste modo, em cada passo da fragmentação, apenas aquelas dimensões com variações significativas da distribuição dos valores são divididas. A uniformidade da distribuição pode ser avaliada empregando um teste estatístico não paramétrico. Dependendo do banco de dados em consideração, a redução da complexidade algorítmica com respeito da quantidade de atributos obtida por este caminho pode tornar desnecessária a inclusão da etapa de *feature selection*. Esta segunda estratégia é uma possível direção para um trabalho futuro.

O algortimo, no seu presente estágio de desenvolvimento, somente considera atributos numéricos lineares. Entretanto, ele pode ser extendido para incluir o caso de atributos categóricos usando abordagens parecidas com a indicada por STANFILL e WALTZ [72]:

a métrica de diferença de valores (*value difference metric*, VDM).

Finalmente, a implementação de uma versão distribuída do nosso algoritmo é uma tarefa para ser considerada no curto prazo.

Autômato celular generalizado como modelo das interações sociais

As relações sociais são atualmente motivo de intensa pesquisa. Os trabalhos se diferenciam uns de outros pelo foco temático e pelo tipo de modelo empregado. Assim por exemplo, há estudos feitos sobre citações de trabalhos científicos, sobre co-participação na produção filmográfica, sobre comunidades virtuais etc. Estes trabalhos geralmente utilizam grafos para modelar a existência de algum tipo de interação; dependendo do caso, o grafo pode ser direcionado – quem referenciou quem –, e suas arestas podem ser ponderadas, sendo que o peso é proporcional à frequência com que as interações acontecem. Em outra linha de pesquisa, as interações são modeladas empregando agentes que se movimentam sobre um reticulado com um esquema de avance que comporta algum tipo de aleatoriedade. Neste tipo de estudo, os próprios agentes são os responsáveis por levar um registro de suas interações. Em principio, quaisquer dois agentes podem se encontrar e interagir durante os seu passeios virtuais; a conectividade entre os elementos é potencialmente total. O autômato celular generalizado que apresentamos neste trabalho constitui um modelo que incorpora caraterísticas das duas abordagens supracitadas. O modelo da conta de representar a estabilidade temporal de algumas interações por meio da componente fixa da vizinhança; simultaneamente, a dinâmica global ganha plasticidade através da componente variável da vizinhança. Ambos os tipos de relações, as estáveis e as esporádicas, devem ser levadas em consideração por aqueles temas de pesquisa que precisem estudar ao mesmo tempo a evolução da dinâmica global e a evolução dos elementos que a compõem. Assim, nosso modelo permite registrar as consequências da relação implícita entre o nível global – macro – e o nível individual – micro.

A regra local de mudança de estado incorpora duas caraterísticas, originais pela sua combinação. A primeira é o uso de um autômato celular diferencial: como fruto da interação com os vizinhos, uma célula determina o sentido da sua eventual mudança de estado; se a pressão for de um teor tal que uma mudança se faz necessária, o novo estado tem como ponto de partida o estado corrente do indivíduo. Destarte é capturada parte da inércia que modula a capacidade para mudar de um indivíduo. Portanto, no nosso modelo, para conhecer o estado dos indivíduos, são levadas em consideração simultaneamente, tanto a conjuntura na qual ele está inserido, como sua própria história individual. Duas escalas de tempo estão envolvidas na evolução do nosso autômato: a primeira, com um certo grau de volatilidade, é a que rege as variações das conjunturas; a segunda, com constantes de tempo demoradas, é a que carateriza a evolução do estado dos indivíduos. As duas escalas de tempo dão conta de dois fenômenos interrelacionados: a formação da

identidade de grupo, ou memória coletiva, representada pela distribuição de frequência dos estados, e a idiossincrasia, expressada pela aptidão particular de cada indivíduo para se adaptar a circunstancias.

A segunda caraterística incorporada na regra local é a presença dos limiares. Eles, junto com a natureza diferencial do autômato, estão na base da formação das idiossincrasias. A acomodação e a sensibilização, mecanismos gerais responsáveis pela manutenção da homeostase nos seres vivos, são modelados na forma como estes limiares são atualizados. Eles conferem plasticidade aos elementos do sistema; destarte, o comportamento dos indivíduos não fica reduzido a um simples reflexo da conjuntura, mas circunstâncias parecidas podem decorrer em respostas totalmente diferentes dependendo da evolução dos limiares. Por este motivo há um ganho na riqueza de trajetórias possíveis para a configuração global do sistema.

O modelo mostrou ter a capacidade de reproduzir o fenômeno observado na distribuição de freqüência das atividades criminosas. Entretanto este resultado é pobre se comparado com a potencialidade do modelo. A primeira linha de trabalho a prosseguir no futuro se refere à metodologia a seguir para fazer a calibração dos parâmetros do modelo: uma relação clara deve ser construída entre a magnitude dos parâmetros e o valor dos observadores estatísticos para casos reais. Ulteriormente, o modelo deve ser confrontado com a tarefa de avaliar as concepções de TARDE e de BANDURA no contexto particular da dinâmica da criminalidade. Ainda que existam modelos que procuram predizer os locais mais prováveis para que um crime acaeça, eles não dão conta de explicar as variações nas taxas de criminalidade. Nesse sentido é importante poder predizer a eficácia de medidas outras que vigiar e castigar, por exemplo, baseadas em estratégias educativas. Não está clara a relação de eficácias de campanhas maciças de difusão com respeito de medidas de menor amplitude, focalizadas em regiões consideradas conflitivas. Estamos persuadidos que a imitação representa a principal ferramenta para modelar o espectro de condutas preferidas por uma população. Nesse sentido, o modelo deve ser acrescido com agentes que atuem como promotores da difusão das mudanças. Na realidade, reconhecemos este tipo de agente nos modelos humanos empregados pela publicidade, os quais servem como referências de comportamento. Entretanto, a publicidade usa preferencialmente uma modalidade inespecífica de influência. Cabe ao modelo avaliar o impacto da distribuição destes agentes entre a população, como modulador dos comportamentos do grupo.

Estreitamente relacionada com as considerações anteriores, uma outra linha de trabalho futuro é o estudo da aparição e instalação daqueles comportamentos comuns aos membros de uma sociedade. Em particular nos referimos a tradições, ou normas sociais de comportamento cujo estabelecimento parece emergir da dinâmica das interações. No ciclo de vida de uma destas pautas sociais deve ser levada em consideração as mudanças na composição da população, consequência do nascimento e da morte de indivíduos, e também devido à existência de fortes correntes migratórias. Interessa avaliar a existên-

cia de uma cadência crítica de reposição da população, por cima da qual o mecanismo de imitação vê comprometida sua eficácia aos efeitos de manter o conjunto de pautas de comportamento caraterísticos de uma sociedade. Esta abordagem para o estudo do ciclo de vida dos hábitos de uma sociedade não utiliza outras considerações que não sejam as relacionadas com a dinâmica de um sistema complexo, e portanto se exime de fazer valorações macroscópicas tipicamente encontrados em trabalhos de sociologia.

Referências Bibliográficas

[1] ARTHUR, D., VASSILVITSKII, S., "k-means++: the advantages of careful seeding", In: *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, (Philadelphia, PA, USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[2] ASUNCION, A., NEWMAN, D., "UCI Machine Learning Repository", 2007.

[3] BAK, P., *How Nature Works: the science of self-organized criticality*. 1 ed. New York, USA, Springer-Verlag, 1996.

[4] BAK, P., TANG, C., WIESENFELD, K., "Self-organized criticality: An explanation of the 1/f noise", *Phys. Rev. Lett.*, v. 59, pp. 381–384, Jul 1987.

[5] BANDURA, A., *Social learning theory*. Emglewood Cliffs - NJ, Prentice Hall, 1977.

[6] BARABÁSI, A. L., *Linked*. 2 ed. New York, USA, Penguim Group, May 2003.

[7] BARABASI, A. L., ALBERT, R., "Emergence of scaling in random networks", *Science*, v. 286, pp. 509 – 512, OCT 15 1999.

[8] BARBOSA, V. C., *Massively Parallel Models of Computation.* Chichester, UK., Ellis Horwood, 1993.

[9] BARREDO, J. I., KASANKO, M., MCCORMICK, N., *et al.*, "Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata", *Landscape and Urban Planning*, v. 64, n. 3, pp. 145 – 160, 2003.

[10] BERKHIN, P., "Survey Of Clustering Data Mining Techniques", tech. rep., Accrue Software, San Jose, CA, 2002.

[11] BEZDEK, J. C., EHRLICH, R., FULL, W., "FCM - The fuzzy c-means clustering algorithm", *Computers & Geociences*, v. 10, n. 2-3, pp. 191 – 203, 1984.

[12] BLEKAS, K., LAGARIS, I. E., "Newtonian clustering: An approach based on molecular dynamics and global optimization", *Pattern Recognition*, v. 40, pp. 1734 – 1744, JUN 2007.

[13] BONABEAU, E., DORIGO, M., THERAULAZ, G., *Swarm Intelligence: From Natural to Artificial Systems*. 1 ed. New York, USA, Oxford University Press, 1999.

[14] CASTRO, P. D., COELHO, G. P., CAETANO, M. F., *et al.*, "Designing ensembles of fuzzy classification systems: An immune-inspired approach", *ARTIFICIAL IMMUNE SYSTEMS, PROCEEDINGS*, v. 3627, pp. 469 – 482, 2005.

[15] CHEN, L., XU, X., CHEN, Y., *et al.*, "A novel ant clustering algorithm based on cellular automata", In: *IAT '04: Proceedings of the Intelligent Agent Technology, IEEE/WIC/ACM International Conference*, (Washington, DC, USA), pp. 148–154, IEEE Computer Society, 2004.

[16] CODD, E. F., *Cellular Automata*. New York, Academic Press, 1968.

[17] COMANICIU, D., MEER, P., "Mean shift: A robust approach toward feature space analysis", *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, v. 24, n. 5, pp. 603 – 619, 2002.

[18] COOK, W., ORMEROD, P., COOPER, E., "Scaling behaviour in the number of criminal acts committed by individuals", *Journal of Statistical Mechanics-Theory and Experiment*, JUL 2004.

[19] CULIK, K., HURD, L. P., YU, S., "Computation Theoretic Aspects of Cellular Automata.", *Physica D*, v. 45, pp. 357 – 378, SEP 1990.

[20] DE CASTRO, L. N., ZUBEN, F. J. V., "An Evolutionary Immune Network for Data Clustering", In: *Proceedings of the IEEE SBRN (Brazilian Symposium on Artificial Neural Networks)*, (Rio de Janerio - Brasil), pp. 84–89, Nov. 2000.

[21] DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., "Maximum Likelihood from Incomplete Data Via em Algorithm", *Journal of the Royal Statistical Society Series B-Methodological*, v. 39, n. 1, pp. 1 – 38, 1977.

[22] DUFFY, J., OCHS, J., "Emergence of money as a medium of exchange: An experimental study", *American Economic Review*, v. 89, pp. 847 – 877, SEP 1999.

[23] FEYNMAN, R., *The Character of Physical Law*. MIT Press, 1965.

[24] FINKEL, R. A., BENTLEY, J. L., "Quad trees a data structure for retrieval on composite keys", *Acta Informatica*, v. 4, pp. 1–9, Mar. 1974.

[25] FORGY, E. W., "Cluster Analysys of Multivariate Data - Efficiency vs Interpretability of Classifications", *Biometrics*, v. 21, n. 3, pp. 768–792, 1965.

[26] HANDL, J., MEYER, B., “Ant-based and swarm-based clustering”, *Swarm Intelligence*, v. 1, pp. 95–113, Dec. 2007.

[27] HANSEN, P., JAUMARD, B., “Cluster analysis and mathematical programming”, *Math. Program.*, v. 79, n. 1-3, pp. 191–215, 1997.

[28] HINNEBURG, A., GABRIEL, H., *Advances in Intelligent Data Analysis VII*, v. 4723/2007 of *Lecture Notes in Computer Science*, ch. DENCLUE 2.0: Fast Clustering Based on Kernel Density Estimation, pp. 70–80. Springer Berlin / Heidelberg, Aug. 2007.

[29] HINNEBURG, A., KEIM, D., “An efficient approach to clustering large multimedia databases with noise”, In: *Proceedings of the 4th ACM SIGKDD*, (New York, NY), pp. 58–65, AAAI Press, 1998.

[30] HINNEBURG, A., KEIM, D., “A general approach to clustering in large databases with noise”, *Knowledge and Information Systems*, v. 5, n. 4, pp. 387–415, 2003.

[31] ILACHINSKI, A., *Cellular Automata A Discrete Universe*. 1 ed. Singapore, World Scientific, 2001.

[32] JAIN, A. K., MURTY, M. N., FLYNN, P. J., “Data clustering: a review”, *ACM Comput. Surv.*, v. 31, n. 3, pp. 264–323, 1999.

[33] JAIN, A., TOPCHY, A., LAW, M., *et al.*, “Landscape of clustering algorithms”, *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, v. 1, pp. 260–263 Vol.1, Aug. 2004.

[34] JAIN, A. K., DUIN, R. P., MAO, J., “Statistical Pattern Recognition: A Review”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 1, pp. 4–37, 2000.

[35] JENSEN, H. J., *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*. 1 ed. Cambridge, United Kingdom, Cambridge University Press, Jan. 1998.

[36] JOHNSON, S., *Emergence: the connected lives of ants, brains, cities, and software*. 1 ed. New York, USA, Scribner, 2004.

[37] JUANICO, D. E., MONTEROLA, C., SALOMA, C., “Allelomimesis as a generic clustering mechanism for interacting agents”, *Physica A-Statistical Mechanics and its Applications*, v. 320, pp. 590 – 600, 2003.

[38] KARYPIS, G., HAN, E. H., KUMAR, V., “Chameleon: Hierarchical clustering using dynamic modeling”, *COMPUTER*, v. 32, n. 8, pp. 68–90, 1999.

[39] KAUFMAN, L., ROUSSEEUW, P. J., *Finding Groups in Data: an Introdution to Cluster Analysis*. John Wiley & Sons, 1990.

[40] KLüVER, J., STOICA, C., SCHMIDT, J., "Formal Models, Social Theory and Computer Simulations: Some Methodical Reflections.", *Journal of Artificial Societies and Social Simulation*, v. 6, n. 2, p. http://jasss.soc.surrey.ac.uk/6/2/8.html, 2003.

[41] KOHONEN, T., "The self-organizing map", *Proceedings of the IEEE*, v. 78, n. 9, pp. 1464 – 1480, 1990.

[42] KOHONEN, T., "The self-organizing map", *NEUROCOMPUTING*, v. 21, n. 1-3, pp. 1 – 6, 1998.

[43] KULLBACK, S., LEIBLER, R. A., "On Information and Sufficiency", *Annals of Mathematical Statistics*, v. 22, n. 1, pp. 79–86, 1951.

[44] LANGTON, C. G., "Computation at the edge of chaos - phase-transitions and emergent computation.", *Physica D*, v. 42, pp. 12 – 37, JUN 1990.

[45] LATANÉ, B., NOWAK, A., LIU, J., "Measuring emergent social phenomena: dynamism, polarization and clustering as order parameters of dynamical social systems", *Behavioral Science*, v. 39, pp. 1–24, 1994.

[46] LINDENMAYER, A., "Mathematical models for cellular interactions in development I. Filaments with one-sided inputs", *Journal of Theoretical Biology*, v. 18, pp. 280–299, Mar. 1968.

[47] LUO, W., CAO, X., WANG, X., "An immune genetic algorithm based on immune regulation", *Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on*, v. 1, pp. 801–806, May 2002.

[48] MACQUEEN, J. B., "Some Methods for classification and Analysis of Multivariate Observations", In: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, pp. 281–297, University of California Press, 1967.

[49] MEILA, M., "Comparing clusterings by the variation of information", *LEARNING THEORY AND KERNEL MACHINES*, v. 2777, pp. 173 – 187, 2003.

[50] MILGRAM, S., "SMALL-WORLD PROBLEM", *PSYCHOLOGY TODAY*, v. 1, n. 1, pp. 61 – 67, 1967.

[51] NG, R. T., HAN, J., "Efficient and Effective Clustering Methods for Spatial Data Mining", In: *Proceedings of the 20th VLDB Conference*, pp. 144–155, 1994.

[52] NOWAK, A., KUS, M., URBANIAK, J., *et al.*, "Simulating the coordination of individual economic decisions", *Physica A*, v. 287, pp. 613 – 630, DEC 1 2000.

[53] NOWAK, A., LATANÉ, B., *Simulating Societies: The computer simulations of social processes*, ch. Simulating the emergence of social order from individual behavior, pp. 63–84. London, UK, University College London Press, 1 ed., 1994.

[54] NOWAK, A., LEWENSTEIN, M., *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, ch. Modeling social change with cellular automata, pp. 249–285. Springer, 1996.

[55] NOWAK, A., SZAMREJ, J., LATANÉ, B., "From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact", *Psychological Review*, v. 97, pp. 362–376, July 1990.

[56] NOWAK, A., VALLACHER, R. R., TESSER, A., *et al.*, "Society of self: The emergence of collective properties in self-structure", *Psychological Review*, v. 107, pp. 39 – 61, JAN 2000.

[57] NOWAK, A., VALLACHER, R. R., ZOCHOWSKI, M., "The emergence of personality: Dynamic foundations of individual variation", *Developmental Review*, v. 25, pp. 351 – 385, SEP-DEC 2005.

[58] NUSSENZVEIG, H. M., *Complexidade & Caos*. 1 ed. Rio de Janeiro, Brasil, Editora UFRJ / COPEA, 1999.

[59] PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U., "Validity index for crisp and fuzzy clusters", *PATTERN RECOGNITION*, v. 37, n. 3, pp. 487 – 501, 2004.

[60] PARSONS, L., HAQUE, E., LIU, H., "Subspace clustering for high dimensional data: a review", *SIGKDD Explor. Newsl.*, v. 6, n. 1, pp. 90–105, 2004.

[61] PELLEG, D., MOORE, A., "Accelerating exact k -means algorithms with geometric reasoning", In: *Knowledge Discovery and Data Mining*, pp. 277–281, 1999.

[62] RASMUSSEN, M., KARYPIS, G., "gCLUTO - An Interactive Clustering, Visualization, and Analysis System", tech. rep.

[63] REDMOND, S. J., HENEGHAN, C., "A method for initialising the K-means clustering algorithm using kd-trees", *Pattern Recognition Letters*, v. 28, n. 8, pp. 965 – 973, 2007.

[64] SAMMON, J. W., "A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers*, v. C 18, n. 5, pp. 401–409, 1969.

[65] SARAFIS, I. A., TRINDER, P. W., ZALZALA, A. M. S., "NOCEA: A rule-based evolutionary algorithm for efficient and effective clustering of massive high-dimensional databases", *APPLIED SOFT COMPUTING*, v. 7, n. 3, pp. 668 – 710, 2007.

[66] SARKAR, P., "A brief history of cellular automats", *ACM Computing Surveys*, v. 32, n. 1, pp. 80 – 107, 2000.

[67] SCHLKOPF, B., SMOLA, A. J., *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.

[68] SHANNON, C., "A Mathematical Theory of Communication", *Bell System Technical Journal*, v. 27, pp. 379–423, 623–656, Oct. 1948.

[69] SHEIKHOLESLAMI, G., CHATTERJEE, S., ZHANG, A., "WaveCluster: A multi-resolution clustering approach for very large spatial databases", In: *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pp. 428–439, 24–27 1998.

[70] SHUAI, D. X., DONG, Y. M., SHUAI, Q., "A new data clustering approach: Generalized cellular automata", *Information Systems*, v. 32, n. 7, pp. 968 – 977, 2007.

[71] SHUAI, D., ZHANG, B., DONG, Y., "Quantum Particles Model for Data Clustering in Enterprise Computing", *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, v. 6, pp. 4602–4607, Oct. 2006.

[72] STANFILL, C., WALTZ, D., "Toward memory-based reasoning", *Commun. ACM*, v. 29, n. 12, pp. 1213–1228, 1986.

[73] TARDE, G., *L'opinion et la foule*. Collection Recherches Politiques, 1 ed. Paris-FR, Les Presses Universitaires, 1989. Original de 1901.

[74] TARDE, G., *Les lois de l'imitation*. 2 ed. Paris - FR, Editions Kimé, 1993. Original de 1890.

[75] THATCHER, J., "Universality in the Von Neumann cellular model.", Tech. Rep. 03105-30-T, University of Michigan, 1964.

[76] TOFFOLI, T., "Cellular Automata as an alternative to (rather than an approximation of) differential-equations in modeling physics.", *Physica D*, v. 10, n. 1-2, pp. 117 – 127, 1984.

[77] TURNER, J., BEGON, M., BOWERS, R. G., "Modelling pathogen transmission: the interrelationship between local and global approaches", *Proceedings of the Royal Society of London Series B-Biological Sciences*, v. 270, pp. 105 – 112, JAN 7 2003.

[78] VON NEUMANN, J., *Theory of Self-reproducing Automata.* Champaign, IL, USA, University of Illinois Press, 1966.

[79] WANG, W., YANG, J., MUNTZ, R. R., "STING: A statistical information grid approach to spatial data mining", In: *Twenty-Third International Conference on Very Large Data Bases* (JARKE, M., CAREY, M. J., DITTRICH, K. R., *et al.*, eds.), (Athens, Greece), pp. 186–195, Morgan Kaufmann, 1997.

[80] WARD, D. P., MURRAY, A. T., PHINN, S. R., "Integrating spatial optimization and cellular automata for evaluating urban change", *Annals of Regional Sciences*, v. 37, pp. 131 – 148, MAR 2003.

[81] WOLFRAM, S., *A New Kind of Science.* Illinois, Wolfram Media, Inc., 2002. Parcialmente disponível em http://www.wolframresearch.com.

[82] WOLFRAM, S., "Statistical mechanics of cellular automata", *Rev. Mod. Phys.*, v. 55, pp. 601–644, Jul 1983.

[83] XAVIER, A., "The Hyperbolic Smoothing Clustering Method", Tech. Rep. 674/05, PESC / COPPE / UFRJ, Apr. 2005.

[84] XU, X. H., CHEN, L., HE, P., "Ant clustering embeded in cellular automata", *Advances in Artificial Life, Proceedings*, v. 3630, pp. 562 – 571, 2005.