LEILA MILFONT RAMEH

ALGORITMO WANG-LANDAU E AGRUPAMENTO DE DADOS SUPERPARAMAGNÉTICO

RECIFE-PE - AGO/2010

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ALGORITMO WANG-LANDAU E AGRUPAMENTO DE DADOS SUPERPARAMAGNÉTICO

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Desenvolvimento de Métodos Estatísticos e Computacionais

Orientador: Prof. Dr. Adauto José Ferreira de Souza

RECIFE-PE - AGO/2010.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ALGORITMO WANG-LANDAU E AGRUPAMENTO DE DADOS SUPERPARAMAGNÉTICO

Leila Milfont Rameh

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 26/08/2010 pela Comissão Examinadora.

Orientador:

Prof. Dr. Adauto José Ferreira de Souza
Universidade Federal Rural de Pernambuco

Banca Examinadora:

Prof. Dr. Borko D. Stosic
Universidade Federal Rural de Pernambuco
DEINFO-UFRPE

Prof. Dr. Tiago Alessandro Espínola Ferreira
Universidade Federal Rural de Pernambuco
DEINFO-UFRPE

Prof. Dr. Francisco George Brady Moreira
Universidade Federal de Pernambuco
DF-UFPE

Dedico este trabalho ao meu irmão Laércio.

Agradecimentos

Listas com nomes sempre me preocupam, desejo não esquecer pessoas importantes, mas há sempre o risco.

Primeiramente gostaria de agradecer ao meu orientador e amigo Adauto, indispensável para a existência deste trabalho e responsável por despertar em mim um certo apreço por física.

Aos professores Borko, Tatjana e Eufrázio pelos ensinamentos.

À Zuleide amigona que vou levar no meu coração como uma das melhores conquistas do mestrado.

A Marco pela presteza e por melhorar sempre o programa.

Aos amigos mestres Andrea, Amanda, Francisco, Kátia, Rejane, Ulisses, Vinícius e principalmente Tadeu que vem comigo de outras batalhas.

Aos amigos que agradeço a Gauss o convívio, em especial a Aranildo, David e Rosilda pelo atendimento home-care. A Paulo, Aranildo e Felipe pelas indicações de leituras. E a Jader pela ajuda com o TEX.

Os cafezinhos com os físicos, Jairo e Paulo, também contribuiram muito para acordar e retomar os trabalhos.

À florzinha Lidiane, irmã em orientação, que me forneceu material e atenção.

Aos meus pais que estavam sempre me lembrando que eu era capaz.

A meu irmão Leo que me faz querer ser melhor para servir de exemplo.

Às minhas irmãs Laedja e Ladjane por estarem comigo sempre. E a Ladjane pelas correções.

Aos meus sobrinhos Gabriel, Rafael e Luís Henrique pelas risadas e momentos doces.

À minha família Rameh que sempre impulsionou em mim uma vontade de crescer.

A Vovô Ide (in memoriam) e vovó Cença, a base sólida.

A tio Toinho e Rodrigo pela contribuição gráfica.

À Rossana por continuar cuidando de mim.

À Marília por estar a disposição para me ajudar ainda que o assunto fosse física.

Ao primo Robson em quem descobri um pai que cuida tão bem de mim.

À tia Nanau sempre ali pronta pra resolver qualquer bronca.

A tio Harry (in memoriam) que continua sendo meu cientista predileto.

À tia Letícia por me fazer acreditar que estava pertinho do fim e por segurar na minha mão pra me ajudar a escrever.

Aos Milfont pelos momentos vitais de descontração.

À tia Tânia por acreditar na próxima etapa acadêmica e me ajudar a ir em busca de um sonho distante.

A tio João por ter me ajudado a escrever e adequar parte do texto.

À Tia Lane, Tahnee e Jensinho porque são simplesmente especiais para mim.

Aos meus enfermeiros de primeira Tati e Gledson.

À Fafá que nos momentos difíceis se fez presente e ainda levava Xandinha e Xandinho pra me animar.

A Luiz Felipe, melhor amigo sempre.

À minha grande amiga Mirella por estar comigo em cada começo, recomeço, fim. À Brenda pelo riso frouxo. À Virgínia por estar presente ainda que fisicamente distante. E a Bruna pelo colo sempre disponível.

À Nicole e Carla por cuidarem sempre de mim.

À Rafaela por acreditar em mim e me dar força.

A Hemílio, Claudyvan, Robinho, Syntia, Flávio, Daniel e Rodrigo por continuarem presentes na minha vida.

A Carlos por ter me ajudado enquanto foi possível e a Pedro por ter me feito muitas vezes esquecer os problemas de gente grande.

Ao amigo fonte bibliográfica George.

À Gisa por se fazer presente na hora certa e da melhor forma possível.

À Tereza, Andrea, Bia e Odael pela amizade e carinho.

A Eduardo pelas ilustrações e por me fazer rir.

A Diogão, sempre que o assunto for programação o primeiro nome que me vem a mente.

A todos que me levantaram daquela queda.

E à CAPES.

There's no remaking reality.
Just take it as it comes.
Hold your ground and take it as it comes.

—PHILIP ROTH (Everyman)

Resumo

O método de agrupamento de dados não supervisionado proposto por Domany e colaboradores baseia-se no mapeamento do problema em um sistema magnético granular não homogêneo, cujas propriedades são investigadas através de algum método de Monte Carlo. A matriz que contém os dados é composta por n atributos de valor numérico e corresponde a um ponto em um espaço euclidiano n-dimensional. A cada item de dado é associado um spin de Potts. A interação entre tais spins decai exponencialmente com o aumento da distância entre eles. Isto favorece o alinhamento dos spins associados a objetos similares. O sistema físico corresponde a um ferromagneto desordenado que, por sua vez, é descrito por um hamiltoniano de Potts de q estados. Espera-se que o sistema magnético exiba três regimes quando sua temperatura seja variada. Para temperaturas muito baixas o sistema está completamente ordenado. No outro extremo, em altas temperaturas, o sistema não apresenta qualquer ordem magnética. Numa faixa intermediária de temperaturas, spins dentro de certas regiões permanecem fortemente acoplados, formando grãos. Porém, um grão não influencie o comportamento de outro grão. Ou seja, os grãos estão não correlacionados. Este estado intermediário caracteriza um estado superparamagnético. A transição de um regime para outro pode ser identificada por picos na curva de calor específico versus temperatura. Aplicamos o método aos conjuntos de dados reais da planta íris e de dados médicos, conhecido por BUPA, aos dados sintéticos conhecidos por Ruspini e a um conjunto de dados, gerado por nós, que consiste de duas figuras tridimensionais sobrepostas, um esfera e um toro. Procedemos a classificação dos dados através da correlação spin-spin em diversas temperaturas. O principal resultado foi a verificação que nem sempre o agrupamento realizado na fase superparamagnética é o ideal.

Palavras-chave: Agrupamento de Dados; Algoritmo de Wang-Landau; Densidade de Estados; Método de Monte Carlo

Abstract

The method of unsupervised data classification proposed by Domany and coworkers is based on mapping the problem onto an inhomogeneous granular magnetic system whose properties can be investigated through some Monte Carlo Method. The array containing the data consists of n numeric attributes corresponding to points in an n-dimensional Euclidean space. Each data item is associated with a Potts spin. The interaction between such spins decays exponentially with the distance. This favors the alignment of the spins associated with similar objects. The physical system corresponds to a disordered ferromagnet which, in turn, is described by a Hamiltonian of a q-states Potts model. It is expected that the magnetic system exhibits three temperature-dependent regimes. For very low temperatures the system is completely ordered. At the other extreme, high temperatures, the system shows no magnetic order. In an intermediate range of temperatures, the spins within certain regions remain tightly coupled, forming grains. However, a grain does not influence the behavior of another grain. That is, the grains are non-correlated and this intermediate state is named a superparamagnetic phase. The transition from one regime to another can be identified by peaks in the specific heat versus temperature curve. We apply the method to several artificial and real-life data sets, such as classification of flowers, summary medical data and identification of images. We measure the spin-spin correlation at several temperatures to classify the data. In disagreement with the Domany and coworkers claims we found that the best classification of the data occurred outside the superparagnetic phase.

Keywords: Data Clustering; Wang-Landau Algorithm; Density of States; Monte Carlo Method

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

O problema de agrupamento de dados consiste em separar ou agrupar objetos de uma dada coleção em classes ou categorias. O objetivo é alocar objetos mais parecidos dentro de um mesmo grupo. É um problema matemático mal colocado, visto que o número de classes pode não ser conhecido de antemão e o termo "parecido" é vago. Em geral, uma medida de similaridade é arbitrada e os objetos são classificados de acordo com o grau de similaridade fornecido por tal medida. Especificamente, dois objetos da mesma classe possuem um grau de similaridade superior ao grau de similaridade atribuído a um par formado por objetos oriundos de classes distintas.

O termo análise de agrupamento é um nome genérico para um conjunto de técnicas e/ou algoritmos dedicados a resolver o problema acima. Uma análise de agrupamento consiste em tentar descobrir estruturas em um conjunto de dados sem no entanto fornecer uma explicação ou interpretação para a existência de relações entre os objetos do conjunto. Logo, análise de agrupamento é uma abordagem exploratória e permite ao pesquisador organizar os dados em alguma estrutura lógica.

Algumas vezes o interesse é classificar os objetos em grupos pré-existentes. Neste caso, a classificação pode ser supervisionada e realizada através de técnicas clássicas multivariadas, por exemplo, análise discriminante [1]. E quando o pesquisador tem algum conhecimento ou pode antecipar relações entre os dados, um modelo paramétrico pode ser empregado. Dentre os métodos supervisionados podemos citar a técnica da função discriminante de Fisher [2], o método do vizinho mais próximo, o método da soma de quadrado dos erros de Ward e o k-médias [3].

Em muitos casos de interesse, porém, não existe um conhecimento sobre a estrutura dos dados ou, ainda, a interferência do pesquisador pode influenciar a qualidade dos agrupamentos. Então, é mais natural adotar aproximações não-paramétricas, que fazem menos suposições sobre o modelo. Dentre os métodos não-supervisionados, entendido como mínima interferência do pesquisador, podemos citar o agrupamento Fuzzy [4], redes neurais [5] e otimização baseada no comportamento de uma colônia de formigas [6, 7], dentre outros. Estes métodos estão

sujeitos, porém, a pelo menos uma das seguintes limitações: alta sensibilidade à inicialização, performance pobre quando os dados contêm agrupamentos sobrepostos, ou inabilidade para manipular variabilidades nas formas dos agrupamentos, densidades dos agrupamentos, e tamanhos dos agrupamentos. Todos estes algoritmos tendem a criar grupos mesmo quando nenhum grupo existe nos dados.

Uma maneira simples de introduzir um problema matemático mal especificado é através de um exemplo. Abaixo, mostramos como o problema de agrupamento pode ser visto como um problema de otimização combinatória. Suponha que dispomos de n objetos rotulados de 1 a n que serão agrupados em k grupos. Existem, então, k^n maneiras distintas de separar os objetos. A distinção entre cada possível particionamento do conjunto pode ser realizada pintando os objetos com k diferentes cores. Por exemplo, para 2 cores e 3 objetos, teremos 8 partições distintas, ver figura (1.1). Caso não conhecêssemos o número de grupos, poderíamos usar desde uma única cor até, no caso mais extremo, n cores. Nesta variante do problema teríamos 27 possíveis partições distintas. Para um conjunto de 3 objetos implicaria em 3^3 possibilidades de agrupamentos como mostra a figura (1.2). Para determinar a partição correta, de acordo com a nossa medida de similaridade previamente arbitrada, necessitamos introduzir uma função objetiva, ou função de mérito. Isto corresponde a atribuir um valor numérico associado a cada um dos particionamentos. Um extremo da função objetiva indica o particionamento ótimo. Infelizmente a abordagem descrita neste exemplo não tem muita utilidade prática. Devido ao crescimento exponencial do número de partições com a cardinalidade do conjunto de objetos esta abordagem direta é computacionalmente inviável.

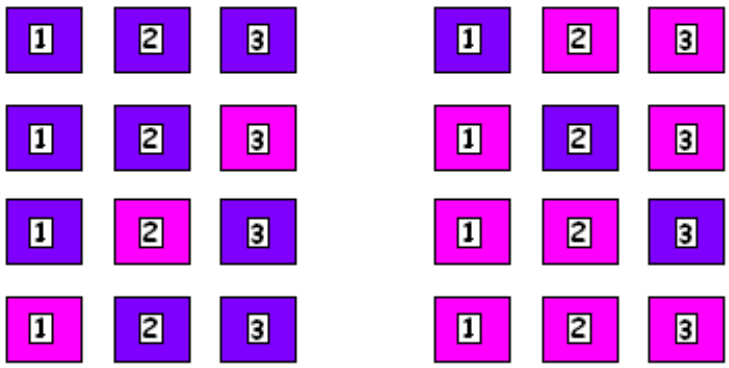

Figura 1.1 Problema de otimização combinatória com o número de grupos conhecido.

Embora trate-se de análise exploratória de dados, o interesse em técnicas de agrupamento de dados surge nas mais diversas áreas, dentre as quais, citaremos segmentação de imagens [8], reconhecimento de objetos tridimensionais [9], recuperação de informações [10] e mineração

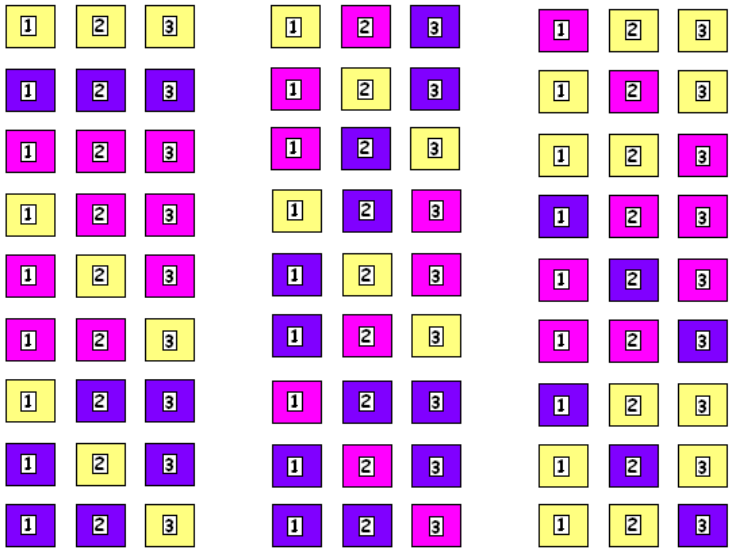

Figura 1.2 Problema de otimização combinatória sem o conhecimento do número de grupos.

de dados [11]. A mineração de dados pode ser exemplificada por análise financeira, análise de DNA [12, 13] e análise de imagens geradas por satélite [12].

A análise de agrupamento de dados está sujeita a algumas arbitrariedades como as escolhas de uma medida de similaridade e de uma função mérito, bem como a informação ou não do número de grupos existente nos dados. Portanto, existe um grande interesse em métodos que evitem tais arbitrariedades. Algumas abordagens inspiradas em sistemas biológicos possuem características atraentes, pois imitam a forma que os organismos vivos resolvem problemas como, por exemplo, otimização baseada no comportamento de uma colônia de formigas [6, 7] e algoritmos genéticos [14].

Nesta dissertação examinaremos detalhadamente o desempenho de um método proposto por Blatt, Wiseman, & Domany [15] baseado nas propriedades físicas de um sistema magnético. A exemplo dos métodos baseados em sistemas biológicos, este método é classificado como um método de agrupamento não-supervisionado. Ou seja, enquadra-se na categoria de heurísticas que “aprendem” a descobrir possíveis estruturas presentes nos dados com o mínimo de interferência do pesquisador. Nesta técnica, o número de grupos não precisa ser previamente conhecido e ela é relativamente insensível à medida de similaridade escolhida. Além disso, não é necessário introduzir uma função de mérito explicitamente. O particionamento dos dados decorre do comportamento termodinamicamente determinado do sistema físico. Uma etapa crucial na aplicação do método é o mapeamento dos dados em um sistema magnético, o qual será detalhado no capítulo 2.

Especificamente, o problema de agrupamento é mapeado no problema de determinar o comportamento termodinâmico de um ferromagneto heterogêneo desordenado. As características dos dados são incorporadas no hamiltoniano que descreve o sistema magnético. Como veremos no capítulo 2, o hamiltoniano corresponde a um modelo de Potts de q estados. Os spins de Potts residem nos nós de uma rede aleatória cuja conectividade depende dos dados em si. A cada nó da rede, portanto a cada spin, está associado um objeto do conjunto de dados. Os spins interagem com seus vizinhos mais próximos com um acoplamento ferromagnético cuja intensidade diminui com a distância. Desta forma, há uma tendência dos spins mais próximos estarem mais fortemente correlacionados. Espera-se que o sistema encontre-se completamente ordenado para temperaturas suficientemente baixas. No outro extremo, temperaturas muito altas, o sistema atinge uma fase paramagnética, na qual os spins estão completamente descorrelacionados. Em temperaturas intermediárias, podem surgir grãos de spins altamente correlacionados entre si, enquanto que diferentes grãos quase não exibem correlação. Logo, os grãos de spins correspondem aos grupos existentes no conjunto de dados. A figura (1.3) representa as fases do sistema magnético correspondente a cada regime de temperatura. Nosso problema é identificar os spins que formam cada grão.

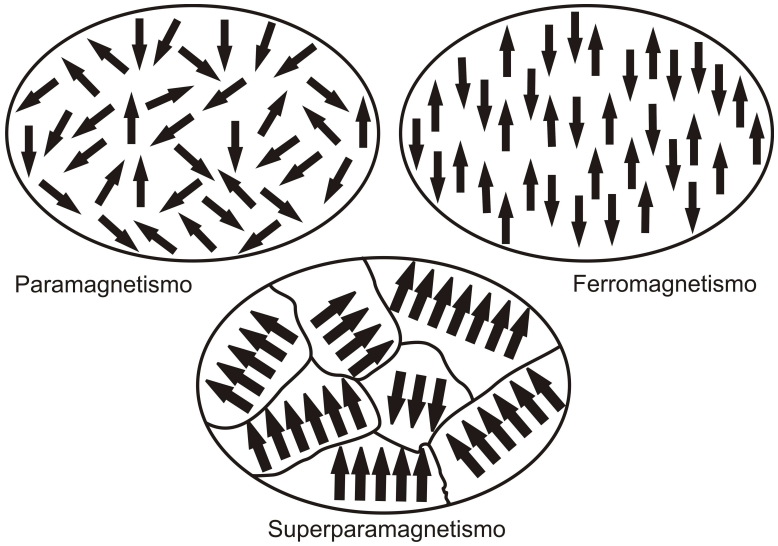

Figura 1.3 Fases do sistema magnético.

Identificar as fases de um sistema magnético desordenado não é um problema trivial. Talvez seja tão difícil quanto o problema original de particionar os dados. Porém, para resolver o segundo problema dispomos de um arsenal de técnicas oriundas da física estatística. Em particular, o método de Monte Carlo [16] é largamente empregado no estudo de modelos de spin na rede [17]. Nesta dissertação utilizaremos uma técnica de Monte Carlo introduzida por Wang

e Landau [18]. O algoritmo Wang-Landau consiste em um passeio aleatório que é executado no espaço de energia para extrair diretamente uma estimativa para a densidade de estados. A probabilidade canônica pode então ser encontrada para qualquer temperatura e as propriedades termodinâmicas podem ser determinadas através de derivadas apropriadas da função de partição. Para uma revisão sobre os tópicos de mecânica estatística ver ([19]).

Nosso objetivo é fundir as duas técnicas promissoras, o agrupamento de dados superparamagnético proposto por Domany et al. [15] e o algoritmo de Wang e Landau [18] para obter a termodinâmica do sistema físico.

No próximo capítulo detalharemos o método de Agrupamento de dados superparamagnético. Discutiremos em detalhes o mapeamento do problema em um modelo de Potts desordenado. No terceiro capítulo, são apresentados o algoritmo de Metropolis [20] e a aproximação do método de Monte Carlo proposto por Wang e Landau [18]. Além disso, discutiremos a amostragem Wang-Landau com intervalo auto-adaptativo [22], que é uma versão do algoritmo de Wang-landau desenvolvida para aplicação em sistemas com uma estrutura da densidade de estados complicada. No quarto capítulo, apresentamos os resultados para diversos bancos de dados. Vamos discutir ainda o desempenho do método e as classificações obtidas. Finalmente, nossas conclusões e perspectivas futuras são apresentadas no quinto capítulo.

CAPÍTULO 2

Agrupamento de dados superparamagnético

O método baseado nas propriedades físicas de um ferromagneto não-homogêneo proposto por Blatt, Wiseman e Domany [15] é uma aproximação para agrupamento de dados. Para a aplicação do método não é necessário fazer suposição a respeito da distribuição de probabilidade inerente aos dados. O método consiste de três estágios. O ponto inicial é a especificação do hamiltoniano que governa o sistema. Em seguida, a localização das temperaturas em que ocorrem as transições de fases do modelo medindo a energia E e o calor específico C como função da temperatura T. Por fim, a medição da correlação G_{ij} entre pares de spins nas faixas de temperatura correspondentes a uma dada fase de equilíbrio do sistema. Esta função de correlação é então usada para particionar os spins e, consequentemente, os correspondentes pontos de dados dentro dos grupos.

Vamos agora associar o conjunto de dados a um ferromagneto heterogêneo desordenado. Assumimos que os dados resultam de N observações e cada uma é rotulada por v_i, com $i = 1, 2, \ldots, N$. Cada v_i possui um conjunto de p atributos e pode ser especificado por uma p-tupla. É conveniente pensar nestas *tuplas* como vetores $\vec{x}_i$ de um espaço métrico p-dimensional. Desta forma, cada item de dado v_i corresponde a um ponto $\vec{x}_i$ em um espaço euclidiano de dimensão apropriada aos dados em questão, como está ilustrado na figura (2.1). Note que, quanto mais próximos os pontos estão, mais similares são os itens correspondentes.

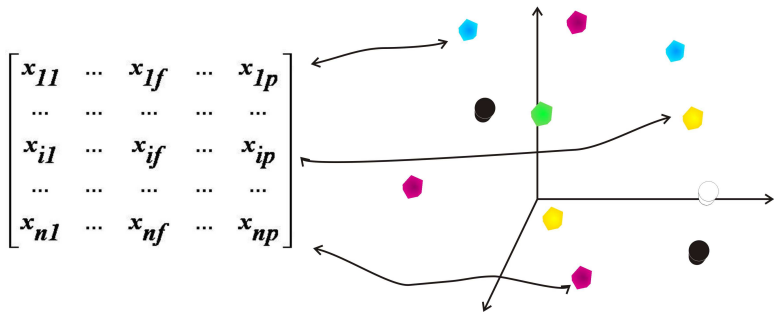

Figura 2.1 Associação dos dados a um espaço euclidiano de dimensão apropriada.

Para realizar o mapeamento do problema de agrupamento de dados em um problema de

spins na rede, fazemos o seguinte. A cada ponto v_i é atribuído uma variável de spin de Potts, ou seja, uma variável que pode assumir um dentre q valores inteiros. Na figura (2.2) ilustramos tal atribuição. Observe que cada objeto possui um "spin" e que na região de baixas temperaturas os spins apontam na mesma direção e é impossível distinguir um objeto a partir do estado do spin a ele associado. À medida que a temperatura aumenta este estado evolui e os spins paulatinamente tornam-se mais independentes. Para temperaturas muito altas cada spin pode apontar com igual probabilidade para uma das q possíveis direções. Como ilustrado na figura, em temperaturas intermediárias os spins associados a objetos de mesmo tipo apontam na mesma direção que resulta a atribuir um super-spin a cada grupo de objetos similares. Em um certo sentido a temperatura funciona como parâmetro que indica a resolução com que analisamos a coleção de objetos. Na ilustração apresentada na figura (2.3) indicamos as interações entre os spins. Observe no destaque que a interação entre objetos do mesmo tipo J_1 e J_2 é menor do que a interação entre objetos distintos, J_3. Ou seja, escolhemos a intensidade da interação como uma função decrescente da distância euclidiana d_{ij} entre os spins.

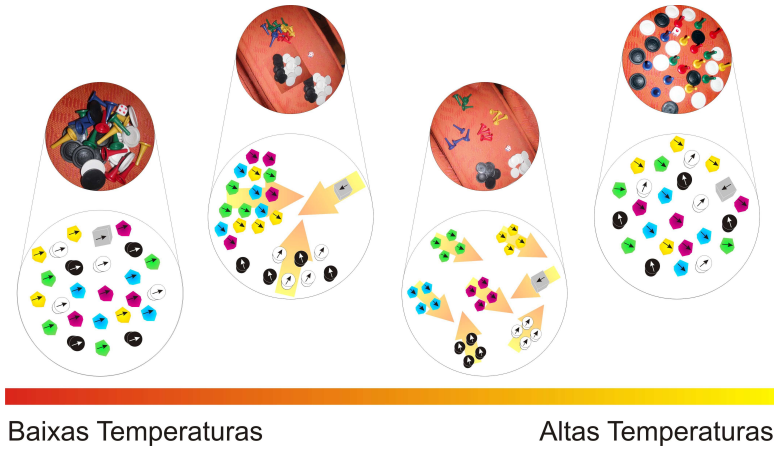

Figura 2.2 Analogia física do problema de agrupamento.

Em um modelo de Potts, a variável de spin s pode assumir um dentre q valores inteiros, $s = 1, 2, \ldots, q$. Escolhemos então o valor de q que determina principalmente a nitidez das transições e as temperaturas em que elas ocorrem. Quanto maior for q, mais nítida será a transição. Porém, com o aumento do valor de q é necessário fazer simulações mais longas, computacionalmente custosas, a fim de manter uma dada acuracidade estatística. O valor de q não implica qualquer suposição sobre o número de grupos presente nos dados. Neste trabalho escolhemos $q = 20$ de acordo com a sugestão de Domany et al. [15]. Uma observação importante é que nenhum dos atributos do hamiltoniano precisa ser ajustado, eles são determinados a partir das características dos dados.

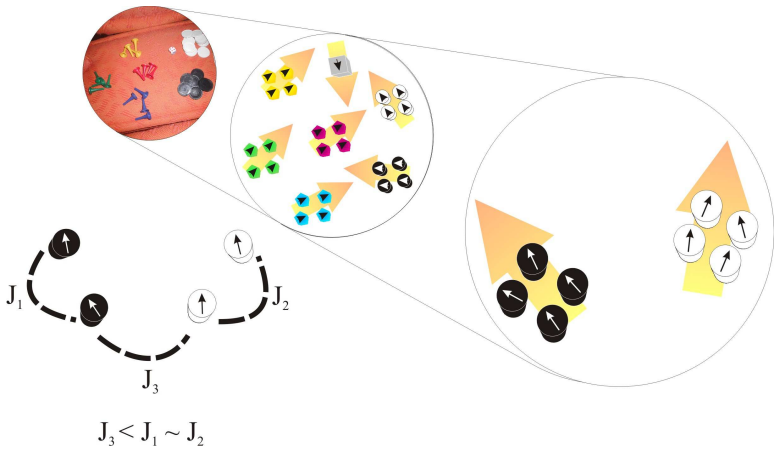

Figura 2.3 Representação da interação entre spins.

Um spin no sítio i interage com um outro spin no sítio j através de um acoplamento ferromagnético de intensidade $J_{ij} > 0$. Denota-se por S uma configuração do sistema, $S = \{s_i\}_{i=1}^{N}$. A energia de uma tal configuração é dada pelo hamiltoniano de Potts

$$\mathscr{H}(S) = \sum_{i,j} J_{ij}(1 - \delta_{s_i,s_j}) \quad s_i = 1, \ldots, q \tag{2.1}$$

em que J_{ij} é a intensidade da interação entre os spins i e j, $i \neq j$ com

$$\delta_{s_i,s_j} = \begin{cases} 1 & \text{se } s_i = s_j \\ 0 & \text{se } s_i \neq s_j. \end{cases} \tag{2.2}$$

A contribuição de um par i e j para $\mathscr{H}$ é 0 quando $s_i = s_j$, isto é, quando os dois spins estão alinhados, e é $J_{ij} > 0$ caso contrário. O hamiltoniano (veja equação (2.1)) é muito similar a outras funções de energia usadas em sistemas neurais, quando cada variável de spin representa um neurônio com uma ligação excitatória entre neurônios interagentes. Para mais detalhes, veja [23].

Para que o hamiltoniano dado pela equação 2.1 reflita as propriedades de um magneto granular fortemente heterogêneo, é necessário que os acoplamentos entre os spins associados aos dados localizados nas regiões de alta densidade sejam fortes, enquanto que os spins correspondentes aos itens de dados que ocupam regiões de baixa densidade interajam apenas fracamente. Para isto, Domany et al [15] introduziram uma escala de comprimento local, a, que fixa uma distância característica a partir da qual a intensidade da interação decai consideravelmente. Tal distância característica é definida pelas regiões de alta densidade e deve ser menor que a distância típica entre os pontos das regiões de baixa densidade.

A intensidade da interação entre dois spins J_{ij} é dada pela equação

$$J_{ij} = \begin{cases} \frac{1}{\hat{K}} \exp\left(-\frac{d_{ij}^2}{2a^2}\right) & \text{se } i \text{ e } j \text{ interagem} \\ 0 & \text{caso contrário} \end{cases} \tag{2.3}$$

A escala de comprimento local, a, é dada pela média de todas as distâncias d_{ij} entre pares i e j de spins interagentes. E $\hat{K}$ é o número médio de interações, isto é, duas vezes o número de interações não-nulas dividido pelo número de pontos N. Com esta escolha, quanto mais próximos dois spins estão maior a chance de eles serem encontrados no mesmo estado, ou seja de eles estarem mais correlacionados.

A fim de calcular a média termodinâmica de uma grandeza física A em uma temperatura fixada T, temos que calcular a soma

$$\langle A \rangle = \sum_S A(S)P(S), \tag{2.4}$$

na qual o fator de Boltzmann,

$$P(S) = \frac{1}{Z} \exp\left(-\frac{\mathscr{H}(S)}{T}\right), \tag{2.5}$$

faz o papel da densidade de probabilidade, que dá o peso estatístico de cada configuração de spin $S = \{s_i\}_{i=1}^N$ em equilíbrio térmico e Z é um fator de normalização dado por

$$Z = \sum_S \exp\left(-\mathscr{H}(S)/T\right). \tag{2.6}$$

Algumas das grandezas físicas mais importantes para este sistema magnético são o parâmetro de ordem ou magnetização e o conjunto de funções δ_{s_i,s_j} porque suas médias térmicas refletem as propriedades de ordem do modelo.

O parâmetro de ordem do sistema é $\langle m \rangle$, em que a magnetização, $m(S)$, associada com uma configuração de spin S é definida como

$$m(S) = \frac{qN_{max}(S) - N}{(q-1)N} \tag{2.7}$$

com

$$N_{max}(S) = max\{N_1(S), N_2(S), \ldots, N_q(S)\}, \tag{2.8}$$

em que $N_\mu(S)$ é o número de spins com o valor μ; $N_\mu(S) = \sum_i \delta_{s_i,\mu}$.

Quando os spins estão em uma rede regular e todos os vizinhos mais próximos interagem com a mesma constante de acoplamento, $J_{ij} = J$, o sistema de Potts é homogêneo. Tais modelos exibem duas fases. A altas temperaturas o sistema é paramagnético ou desordenado, $\langle m \rangle = 0$, indicando que $N_{max}(S) \approx N/q$ para todas as configurações estatisticamente significantes. Nesta fase a função de correlação G_{ij} decai para $1/q$ quando a distância entre os pontos i e j é grande. Esta é a probabilidade de encontrar dois spins de Potts completamente independentes no mesmo estado. A temperaturas muito altas sítios vizinhos têm igualmente $G_{ij} \approx 1/q$.

Conforme a temperatura é reduzida, o sistema sofre uma forte transição para uma fase ordenada ferromagnética, a magnetização salta para $\langle m \rangle \neq 0$. Isto significa que nas configurações fisicamente relevantes, a baixas temperaturas, um estado de Potts domina e $N_{max}(S)$ excede N/q por um número macroscópico de sítios. A temperaturas muito baixas $\langle m \rangle \approx 1$ e $G_{ij} \approx 1$ para todos os pares i e j.

A suscetibilidade é uma medida da flutuação da magnetização

$$\chi = \frac{N}{T}\left(\langle m^2 \rangle - \langle m \rangle^2\right). \tag{2.9}$$

que também reflete as fases termodinâmicas do sistema. A baixas temperaturas, flutuações da magnetização são desprezíveis, então a suscetibilidade χ é pequena na fase ferromagnética.

Considere a situação em que os spins formam "grãos" magnéticos. Neste caso, as ligações entre vizinhos que pertencem ao mesmo grão são muito fortes e entre todos os outros pares muito fracas. Tal situação pode ser modelada por um modelo de Potts fortemente não homogêneo. A temperaturas baixas, este sistema é também ferromagnético, mas como a temperatura está aumentando, o sistema pode exibir uma fase intermediária superparamagnética. Nesta fase, grãos fortemente acoplados estão alinhados, isto é, estão em suas respectivas fases ferromagnéticas, enquanto não existe ordem relativa em diferentes grãos.

Na temperatura de transição da fase ferromagnética para a superparamagnética é observado um nítido pico de χ. Na fase superparamagnética, flutuações do estado tomado pelos grãos que atuam como um todo, isto é, como gigantes superspins, produzem grandes flutuações na magnetização. Com a temperatura aumentando demais, a transição superparamagnética-paramagnética é alcançada. Cada grão desordena-se, e χ é reduzido abruptamente através de um fator que é aproximadamente do tamanho do maior grupo. Portanto, as temperaturas entre dois picos na suscetibilidade indicam o intervalo de temperaturas em que o sistema está em uma das suas fases superparamagnéticas.

Como a temperatura está aumentando, o sistema pode se separar primeiro em dois grupos, cada um dos quais se quebrando em mais subgrupos macroscópicos e assim por diante. Tal

estrutura de hierarquia dos agrupamentos magnéticos reflete uma organização hierárquica dos dados dentro de categorias e subcategorias.

Os itens de dados do nosso problema de agrupamento serão usados como sítios de um ferromagneto de Potts não homogêneo. A presença de grupos nos dados origina grãos magnéticos do tipo descrito acima nos correpondentes modelos de Potts. Trabalhando na fase superparamagnética do modelo, os valores da função de correlação dos pares de spin de Potts são usados para decidir se um par de spins está ou não dentro do mesmo grão, e associar estes grãos aos grupos de nossos dados.

A necessidade de identificação dos vizinhos de um ponto $\vec{x}_i$ seria eliminada fazendo todos os pares $i,\ j$ de spins de Potts interagirem entre si através de uma interação de curto alcance $J_{ij} = f(d_{ij})$. Esta interação decai exponencialmente com a distância entre os dois pontos de dados. As propriedades das fases e agrupamentos do modelo não serão afetadas fortemente pela escolha de f. Tais modelos têm $O(N^2)$ interações, que tornam suas simulações muito caras para N grande. Por conveniência computacional, mantemos apenas as interações de um spin com um número limitado de vizinhos e ajustamos todos os outros J_{ij} para zero. Uma vez que os dados não formam um reticulado regular, construímos um grafo completamente conectado para definir a rede de interação entre os spins da forma descrita abaixo.

Para dois spins v_i e v_j serem vizinhos, ou interagentes, é necessário que v_i pertença a K-vizinhança de v_j e, vice-versa, v_j pertença a K-vizinhança de v_i. A K-vizinhança de um dado v_i é formada pelos K pontos mais próximos de v_i. Escolhemos K tal que as interações liguem todos os pontos de dados em um grafo completamente conectado. Claramente, K cresce com a dimensionalidade, por isso, achamos conveniente, escolher K de acordo com a dimensionalidade dos dados.

Ao trabalhar com bases de dados multidimensionais comumente nos deparamos com diferença na natureza dos dados que pode ser acompanhada por grandes diferenças nas escalas das variáveis. Usualmente as variáveis são padronizadas antes das distâncias serem calculadas, de modo que todas as variáveis sejam igualmente importantes na determinação das distâncias entre os objetos.

A importância da padronização pode ser exemplificada por um estudo em que o objetivo é classificar árvores de acordo com sua idade, considerando os atributos altura da árvore e diâmetro à altura do peito. Tanto árvores novas quanto árvores velhas podem apresentar um valor elevado de altura, porém o diâmetro das árvores mais velhas tende a ser maior. Ao calcular as distâncias entre as árvores, o diâmetro seria desprezado por conta da pequena escala e apenas os valores das alturas contribuiriam significativamente para as distâncias. E desta forma a classificação seria determinada apenas pela altura das árvores.

A necessidade de padronização dos dados torna-se ainda mais importante para a realização do agrupamento de dados através do método em estudo, pois as interações entre os spins são função da distância entre os objetos e desejamos que todos os atributos contribuam igualmente para a distância entre os spins.

Portanto, sempre que se fez necessário, tomamos o cuidado de padronizar os dados dos conjuntos estudados. Dentre as várias opções de padronização utilizamos a dada pela função:

$$y_{if} = \frac{x_{if} - x_{fmin}}{x_{fmax} - x_{fmin}}, \tag{2.10}$$

em que x_{if} é o f-ésimo atributo do i-ésimo dado não padronizado e x_{fmin} e x_{fmax} são o menor e o maior valor, respectivamente, do i-ésimo atributo na base de dados. Não padronizamos os dados referentes a objetos geométricos, pois seus atributos foram medidos na mesma escala.

Segundo Domany et al. [15] os vários intervalos de temperatura em que o sistema auto-organiza os agrupamentos dentro de diferentes partições podem ser identificados medindo a suscetibilidade χ como uma função da temperatura. Um método de Monte Carlo é aplicado em uma faixa de temperaturas e é obtida uma estimativa da maior temperatura de transição para o regime superparamagnético. Iniciando a partir desta estimativa, podemos tomar varreduras de temperaturas cada vez mais refinadas e calcular a função $\chi(T)$ por simulação de Monte Carlo. Em contraste, através do algoritmo Wang-Landau [18] em uma única varredura calculamos a densidade de estados com a qual obtemos a termodinâmica do sistema para qualquer temperatura.

A fase superparamagnética pode conter muitas diferentes subfases com diferentes propriedades de ordem. Um exemplo típico pode ser gerado pelos dados com uma estrutura hierárquica, dando origem a diferentes partições dos dados aceitáveis. Domany et al. [15] mediram a suscetibilidade χ em diferentes temperaturas a fim de localizar estes diferentes regimes. O objetivo é identificar as temperaturas em que o sistema muda sua estrutura. Em nosso estudo calculamos o calor específico para todas as temperaturas e identificamos as mudanças no sistema físico através dos picos no gráfico do calor específico como função da temperatura.

Os picos nas curvas do calor específico versus temperatura assinalam transições entre as diversas fases do sitema. Nas temperaturas dos picos um grande grupo quebra dentro de grupos um pouco menores, mas ainda macroscópicos. O pico correspondente a temperatura mais alta assinala uma transição superparamagnética-paramagnética, em que um ou mais grupos grandes derretem-se, isto é, quebram-se em muitos grupos menores.

Uma vez que a fase superparamagnética e suas diferentes subfases tenham sido identificadas, selecionamos uma temperatura em cada região de interesse. A razão para tal escolha é que

cada subfase caracteriza um particular tipo de partição dos dados, com novos grupos fundindo ou quebrando. Por outro lado, como a temperatura varia dentro de uma fase, espera-se apenas a diminuição ou expansão dos grupos existentes, mudando apenas a classificação dos pontos nas bordas dos grupos.

Após identificar as temperaturas em que ocorrem as transições de fase, verificaremos a correlação spin-spin. A correlação spin-spin é modificada cada vez que um spin muda de estado, uma vez que o processo é dinâmico. Porém o interesse é verificar a correlação na faixa de temperatura entre a penúltima e a última transições de fase, ou seja antes do sistema degenerar. Para o modelo de Potts, definimos a média térmica de δ_{s_i,s_j}, chamada de função de correlação G_{ij} ou correlação spin-spin, por

$$G_{ij} = \langle \delta_{s_i,s_j} \rangle \tag{2.11}$$

que é a probabilidade de dois spins estarem alinhados. Observe que, no nosso caso, i e j são pares de sítios vizinhos, ou seja, $J_{ij} \neq 0$.

Dada uma configuração de spins, gerada pela simulação, essa função é obtida da seguinte forma. Para cada par de spins separados por uma distância d_{ij}, somamos $+1$ se os dois spins estão alinhados e 0 se estiverem desalinhados.

Usamos a função de correlação G_{ij}, entre sítios vizinhos v_i e v_j, para construir os grupos de dados. A princípio, temos que calcular a média térmica de δ_{s_i,s_j} a fim de obter G_{ij} através de um método de Monte Carlo que será discutido no próximo capítulo. Consideramos que dois spins estão correlacionados se $G_{ij} > 0,5$.

Os grupos dos dados são identificados em três passos:

1. Construímos os núcleos dos grupos usando um método limiar, se $G_{ij} > 0,5$, uma ligação é definida entre os pontos de dados vizinhos v_i e v_j. O grafo conectado resultante depende fracamente do valor usado neste limiar (0,5), desde que ele é maior que $1/q$ e menor que $1-2/q$. A razão é que a distribuição de G_{ij} atinge picos intensos nestes dois valores e é muito menor entre eles.

2. Capturamos pontos situados na periferia dos grupos ligando cada ponto v_i aos seus vizinhos v_j de correlação máxima G_{ij}. Pode acontecer, certamente, que pontos v_i e v_j já tenham sido ligados no passo anterior.

3. Grupos de dados são identificados como componentes ligados nos grafos obtidos nos passos 1 e 2.

Em síntese, apresentamos a seguir o esboço dos estágios e subestágios do algoritmo.

- Construir a analogia física do problema com o modelo de spin de Potts:
  - Associar uma variável spin de Potts $s_i = 1, 2, \ldots, q$ para cada ponto v_i;
  - Identificar os vizinhos de cada ponto v_i de acordo com o critério de vizinhança mútua;
  - Calcular a interação J_{ij} entre pontos vizinhos v_i e v_j.

- Localizar a fase superparamagnética:
  - Estimar a energia média para diferentes temperaturas;
  - Usar o calor específico para localizar a fase superparamagnética.

- Voltar aos dados:
  - Medir a correlação spin-spin, G_{ij}, na fase superparamagnética, para todos os pontos vizinhos v_i, v_j;
  - Construir os grupos de dados.

Para resolver o sistema físico no qual o problema foi mapeado, o comportamento da magnetização em função da temperatura e a função de correlação spin-spin em uma dada temperatura devem ser estimados e usualmente é utilizado algum método de Monte Carlo.

CAPÍTULO 3

Método de Monte Carlo

3.1 Algoritmo Metropolis

Avaliação direta de somas como a da equação (2.4) é impraticável, uma vez que o número de configurações aumenta exponencialmente com o tamanho N do sistema. Métodos de simulação de Monte Carlo superam este problema gerando um subconjunto de configurações característico, que é usado como uma amostra estatística. Eles são baseados na noção de amostragem por importância, em que um conjunto de configurações de spins $s_1, s_2, \ldots, s_M$ é gerado de acordo com a distribuição de probabilidade de Boltzmann (veja equação (2.5)). Então, a equação (2.4) é reduzida a uma média aritmética simples,

$$\langle A \rangle \approx \frac{1}{M} \sum_{i}^{M} A_{(s_i)}, \tag{3.1}$$

na qual o número de configurações na amostra, M, é muito menor que q^N, o número total de configurações. A estimativa será tanto melhor quanto maior for o valor de M.

Um dos algoritmos mais utilizados para amostrar as configurações aleatoriamente é o algoritmo de Metropolis que gera a distribuição de probabilidade de Boltzmann [20]. O algoritmo permite essencialmente que configurações mais prováveis tenham mais chance de ocorrer na média das grandezas de interesse. O algoritmo pode ser descrito como segue.

- Escolhemos uma configuração arbitrária s_i, $i = 0$.
- Sorteamos uma nova configuração s_{i+1} aleatoriamente.
- Calculamos $\Delta E = E_{s_{i+1}} - E_{s_i}$, em que a energia E é dada pelo hamiltoniano descrito na equação (2.1).
  - Se $\Delta E \leq 0$, então a mudança é aceita. Ou seja, se o sistema vai para uma nova configuração mais provável ela é aceita.

– Se $\Delta E > 0$, a mudança pode ainda ser aceita com uma probabilidade $p = \exp(-\beta\Delta E)$. Geramos um número aleatório ε uniformemente distribuído no intervalo $[0,1]$, se $\varepsilon \leq p$ a mudança é aceita, caso contrário uma nova configuração é sorteada aleatoriamente e o processo é repetido.

O algoritmo de Metropolis possibilita, assim, que se procure por configurações estacionárias que são energeticamente favoráveis para os sistemas físicos. Como a configuração inicial é arbitrária, descartamos os primeiros passos para que o sistema se equilibre.

3.2 Algoritmo Wang-Landau

Enquanto o número total de configurações cresce exponencialmente com o tamanho do sistema, o número total de níveis de energia possíveis cresce linearmente com o tamanho do sistema, então é menos custoso calcular $g(E)$ com um passeio aleatório no espaço de energia para um sistema grande. Por exemplo, para um modelo de Potts com q estados, em um quadrado LxL com interações do vizinho mais próximo, o número de níveis de energia possíveis para $q \geq 3$ é cerca de $2N$, em que $N = L^2$ é o número total de sítios no quadrado. Em contraste, o número médio de possíveis estados para cada nível de energia é maior que $q^N/2N$, em que q^N é o número total de possíveis configurações do sistema. Claramente, computadores atuais não são capazes de realizar todos os estados possíveis para calcular qualquer grandeza termodinâmica ainda que a maioria dos modelos em física estatística seja bem definida. É por isso que algoritmos de simulação eficiente e rápida são requeridos nas investigações numéricas.

O método de simulação de Monte Carlo, dito padrão, é flexível e de fácil implementação. Porém, próximo às transições de fase este método apresenta ineficiências devido às grandes flutuações e correlações espaço-temporais que se desenvolvem. Por exemplo, barreiras de energia livre separando estados metaestáveis de estáveis de transições de fase de primeira ordem podem evitar que todo espaço de regiões de configurações importantes seja amostrado. Problemas similares surgem devido ao perfil de energia áspera ou desaceleração crítica próxima às transições de fase de segunda ordem. Muitos algoritmos mais eficientes têm sido propostos para contornar este problema, mas eles são um tanto limitados em aplicabilidade e não calculam a densidade de estados, $g(E)$, corretamente para sistemas grandes. O algoritmo de Wang-Landau [18] oferece vantagens substanciais sobre as aproximações de Monte Carlo existentes.

Uma das características mais atraentes do algoritmo de Wang-Landau é que mesmo para determinar densidades de estado com picos acentuados, quase nenhum conhecimento da fun-

ção é requerido. Inicia-se a simulação com uma densidade plana, ou seja, $g(E) = 1$, que automaticamente se adapta e se molda com acuracidade cada vez maior durante cada passo de Wang-Landau.

Para extrair o comportamento coletivo (macroscópico) de um sistema de partículas interagentes é necessário realizar a mecânica estatística do sistema. Uma prescrição consiste em obter a função de partição canônica

$$Z = \sum_{\{\text{configurações}\}} e^{-\beta E} = \sum_{E} g(E) e^{-\beta E} \tag{3.2}$$

em que $\beta = 1/(k_B T)$. A energia E é dada pelo hamiltoniano da equação (2.1).

A densidade de estados permite calcular em todas as regiões de temperatura sem simulações múltiplas a maioria das grandezas termodinâmicas tais como energia livre e entropia, que não são diretamente avaliadas a partir de simulações de Monte Carlo convencionais. A conexão com a termodinâmica é obtida através da energia livre de Helmholtz

$$F = -kT \log(Z) \tag{3.3}$$

a partir da qual as grandezas termodinâmicas são calculadas tomando-se derivadas apropriadas.

Outras propriedades termodinâmicas podem também ser calculadas a partir da densidade de estados, por exemplo, a energia interna é dada por:

$$U(T) = \frac{\sum_E E g(E) e^{-\beta E}}{\sum_E g(E) e^{-\beta E}} = \langle E \rangle_T \tag{3.4}$$

e o calor específico pode ser estimado a partir das flutuações na energia interna:

$$C(T) = \frac{\partial U(T)}{\partial T} = \frac{\langle E^2 \rangle_T - \langle E \rangle_T^2}{T^2} \tag{3.5}$$

Portanto, uma única simulação poderia ser suficiente para fornecer todas as informações interessantes sobre o sistema.

A densidade de estados $g(E)$, porém, é usualmente desconhecida. De fato, este é o objetivo do Método de Wang-Landau ao calcular esta função e tal objetivo é alcançado da seguinte forma. Primeiro, a densidade de estados é iniciada com um chute, digamos, $g(E) = 1$ para todas as energias E possíveis. Então, um passeio aleatório no espaço de energia é iniciado modificando spins aleatoriamente com a probabilidade para um dado nível de energia proporcional a $1/g(E)$. Se E_1 e E_2 são as energias antes e após a mudança do estado de um spin, a

probabilidade de transição de um nível de energia E_1 para um E_2 é:

$$p(E_1 \rightarrow E_2) = min(\frac{g(E_1)}{g(E_2)}, 1) \tag{3.6}$$

Ou seja, se $g(E_1) \geq g(E_2)$ a transição de E_1 para E_2 é aceita e se $g(E_1) < g(E_2)$ o sistema pode continuar em E_1 ou passar para E_2 com probabilidade $g(E_1)/g(E_2)$.

Cada vez que um nível de energia E é visitado, o valor existente é multiplicado pelo fator de modificação $f > 1$, isto é, $g(E) \rightarrow g(E) * f$, tal que torna-se menos provável visitar estados com a mesma energia novamente. Devido a este ajustamento iterativo do peso de amostragem, todas as energias em um dado intervalo são geradas aproximadamente com a mesma probabilidade. Mesmo quando o passeio aleatório rejeita a mudança, isto é, a energia permanece inalterada, $g(E)$ é modificado com o mesmo fator. Uma razoável escolha do fator de modificação inicial é $f = f_0 = e^1 = 2,71828\ldots$ que permite $g(E)$ se desenvolver rapidamente.

Durante o passeio aleatório, o histograma $H(E)$, o número de visitas para cada nível de energia E, é acumulado. Quando o histograma é aproximadamente plano, a densidade de estados terá convergido para o valor verdadeiro com uma acuracidade proporcional ao fator de modificação $\ln(f)$. O fator de modificação é então reduzido para, por exemplo, $f_1 = \sqrt{f_0}$, o histograma é zerado, e um novo passeio aleatório é iniciado. Este método iterativo continua até que o fator de modificação seja menor que um valor pré-definido, por exemplo, $f_{final} = \exp(10^{-8}) \simeq 1,00000001$).

O fator de modificação atua como um parâmetro de controle para a acuracidade de $g(E)$ durante a simulação e também determina quantas varreduras de Monte Carlo são necessárias para a simulação inteira. É impossível obter um histograma perfeitamente plano e o termo "histograma plano" quer dizer que todas as entradas do histograma não são menores que x% da média do histograma $\langle H(E) \rangle$, em que x% é escolhido de acordo com o tamanho e complexidade do sistema e a acuracidade desejada de $g(E)$.

Note que o algoritmo não satisfaz exatamente a condição de balanceamento detalhado para estágios iniciais da iteração já que $g(E)$ é modificada constantemente durante o passeio aleatório. Porém, ele rapidamente converge para o verdadeiro valor depois de algumas iterações e $f \rightarrow 1$. Se $p(E_1 \rightarrow E_2)$ é a probabilidade de transição do nível de energia E_1 para o nível E_2, a partir da equação (3.2), a razão das probabilidades de transição de E_1 para E_2 e de E_2 para E_1 é:

$$\frac{p(E_1 \rightarrow E_2)}{p(E_2 \rightarrow E_1)} = \frac{g(E_1)}{g(E_2)} \tag{3.7}$$

em que $g(E)$ é a densidade de estados. Em outras palavras, o algoritmo do passeio aleatório satisfaz a condição de balanceamento detalhado:

$$\frac{1}{g(E_1)} p(E_1 \to E_2) = \frac{1}{g(E_2)} p(E_2 \to E_1) \tag{3.8}$$

em que $1/g(E_1)$ é a probabilidade para o nível de energia E_1 e $p(E_1 \to E_2)$ é a probabilidade de transição de E_1 para E_2 para o passeio aleatório. Por fim, a condição de balanceamento detalhado é satisfeita com acuracidade proporcional ao fator de modificação $\ln(f)$.

Podemos resumir o algoritmo de Wang-Landau da seguinte forma.

- sortear aleatoriamente uma configuração inicial
- calcular a energia dessa configuração
- sortear um spin e tentar invertê-lo
- calcular o custo energético associado a tal inversão
  - ΔE
- calcular a energia associada a essa nova configuração
- verificar se o sistema muda para essa nova configuração
  - $p(E_1 \to E_2) = min\left(\frac{g(E_1)}{g(E_2)}, 1\right)$
    - se $g(E_1) \geq g(E_2)$: passa de E_1 para E_2
    - se $g(E_1) < g(E_2)$ pode continuar em E_1 ou passar para E_2 com probabilidade $\frac{g(E_1)}{g(E_2)}$
- atualizar $g(E)$ para a configuração escolhida
- atualizar $H(E)$
- atualizar f

3.3 Amostragem Wang-Landau com intervalo auto-adaptativo

Uma versão de auto-adaptação do algoritmo de Wang-Landau, que é idealmente apropriada para aplicação em sistemas com uma estrutura complicada da densidade de estados, foi proposta por Tröster e Dellago [22]. Para produzir um histograma plano é necessário especificar o intervalo de amostragem no espaço de energia e certificar que todas as energias nesta faixa de configurações acessíveis existem. De fato, se o sistema é proibido de visitar certos valores de energia com a faixa especificada, o histograma de energia correspondente não se tornará plano e o algoritmo não convergirá.

A primeira vista, isto pode ser considerado como uma questão trivial. Poder-se-ia, por exemplo, resolver o problema de energias inacessíveis dividindo toda faixa de energia em muitas janelas sobrepostas e realizar separadamente uma simulação de Wang-Landau em cada uma destas janelas. Possíveis energias inacessíveis ocorrem com maior probabilidade nos limites superior e inferior da faixa de energia (ver figura (3.1)). Portanto, janelas correspondentes a tais energias seriam descartadas e a densidade de estados calculada apenas a partir das janelas onde há convergência.

No caso de modelos contínuos, o problema de possíveis furos pode ocorrer, porém, em sistemas discretos espera-se que isso ocorra como uma regra. Problemas graves também podem surgir se as densidades são calculadas como uma função de mais de uma variável. Considerando o campo magnético H a soma sobre as configurações na equação (3.2) pode ser convenientemente escrita como

$$Z = \sum_{E,M} g(E,M) \exp(-\beta \mathscr{H}) \tag{3.9}$$

em que $g(E,M)$ é o número de configurações microscópicas do sistema com energia E e magnetização M. A função $g(E,M)$ é conhecida como densidade de estados.

Não procedemos ao cálculo de $g(E,M)$ pois o histograma envolveria uma dimensão superior ao histograma para $g(E)$ o que tornaria mais complexa a implementação e $g(E)$ é suficiente para realizar os agrupamentos através do cálculo do calor específico. Utilizamos a proposta para determinação do espaço acessível de energia e magnetização o que torna mais eficiente o método Wang-Landau, pois evita que configurações não sejam visitadas e que o passeio aleatório fique armadilhado em alguma configuração inexistente. Apesar de não trabalharmos diretamente com a magnetização, determinamos o espaço acessível de energia e magnetização porque é mais simples determinar os limites inferior e superior da magnetização que envolve

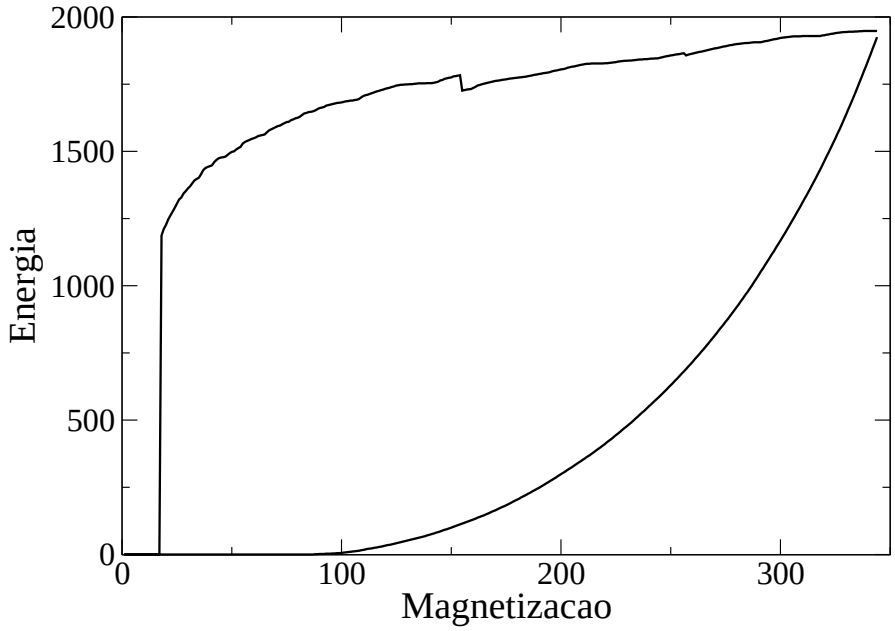

Figura 3.1 Limites inferior e superior da energia e magnetização para a base de dados BUPA [21].

apenas soma de variáveis do que da energia que envolve multiplicação de variáveis.

A seguir será dada uma descrição do algoritmo em que os limites de E e M são determinados adaptativamente durante a simulação. A fim de criar um histograma multidimensional, primeiro obtemos os limites superior e inferior de E e M. Confinamos os parâmetros em uma região retangular com os limites inferior e superior de magnetização denotados por (E_L, M_L) e (E_H, M_H), respectivamente.

Em seguida, escolhemos as larguras do histograma para ambos E e M. n_E e n_M denotam os números de caixas do histograma de largura $(E_H - E_L)/n_E$ e $(M_H - M_L)/n_M$, respectivamente. Os valores centrais da caixa (i, j) são denotados por (E_i, M_j). Nesta grade no espaço $E - M$, são criados arranjos bi-dimensionais para as densidades de estado, $g(E, M)$, e para a frequência $h(E, M)$ para cada par (E, M).

A ideia fundamental do algoritmo de Wang-Landau modificado é registrar sucessivamente os valores mínimo e máximo absolutos $E_{min}(M_j)$ e $E_{max}(M_j)$ para cada índice de magnetização M_j em cada passo do passeio aleatório. Se inicializarmos o passeio aleatório escolhendo uma configuração de spin aleatória com energia na caixa centrada em E_k e a magnetização na caixa centrada em M_l, estes valores servirão então como valores iniciais $E_{max}(M_l) = E_{min}(M_l) = E_k$. Sempre que durante a simulação uma nova magnetização ocorre, este método de inicialização

é realizado para esta particular magnetização. Os limites de energia $E_{max}(M_j)$ e $E_{min}(M_j)$ são atualizados toda vez que uma energia E fora do atual limite de energia pertecente a M_j é encontrada. Para todos os valores de energia entre os limites de energia anterior e atual o histograma correspondente e as entradas da densidade de estados $H(i, j)$ e $g(i, j)$ são inicializados com 0 e 1, respectivamente. Por iteração deste método, o domínio da simulação de Wang-Landau rapidamente se espalha no espaço paramétrico. O critério do histograma plano deve ser checado após a realização de um número elevado, por exemplo, 10^6 passos aleatórios para evitar uma parada prematura desta primeira amostragem do domínio. Ao longo da simulação, o histograma $H(i, j)$ satisfaz um certo critério de histograma plano exceto para pontos fora dos limites de certas áreas isoladas ("buracos") dentro desta região, onde as entradas do histograma são estritamente zero. Como uma segunda condição, insistimos que cada caixa do histograma seja visitada no mínimo, por exemplo, 10^6 vezes. Para identificar a fronteira e buracos do domínio do parâmetro, chamaremos de um "buraco" uma caixa (i, j) encontrada dentro da região do parâmetro que tenha apenas entradas zero no histograma, enquanto o restante do histograma é plano e permanece assim por um grande número de escolhas adicionais de movimentos aleatórios.

Uma vez que os buracos e as caixas fora do domínio de parâmetros foram determinados, o algoritmo de Wang-Landau padrão é realizado simplesmente ignorando essas caixas na verificação do critério do histograma plano, o que equivale a aproximar $g(E,M)$ por $g(E,M) \approx 0$ para os correspondentes valores (E,M). Quando o passo de Wang-Landau é bem sucedido, a região de parâmetro encontrada nesta amostragem do domínio é congelada.

CAPÍTULO 4

Resultados e Discussões

Para desenvolver o método de agrupamento proposto é necessário obter uma rede de vizinhança que descreva os dados. Para isso, definimos um número médio de vizinhos $\hat{K}$ e identificamos os K vizinhos de cada item. Em seguida verificamos quais as vizinhanças que serão confirmadas de acordo com o critério de vizinho mútuo. Em seguida verificamos se há algum item que tenha ficado desconectado, se houver, ele receberá como vizinho o item que estiver mais próximo dele. Portanto o número de vizinhos não será exatamente K, será um número suficiente para gerar uma rede completamente conectada de vizinhos-mútuos.

Para verificar as configurações de energia e magnetização acessíveis procedemos ao método proposto por Trüster e Delago. Varremos o espaço de energia e magnetização com critério de parada $f = 0,001$ e $\tau = 100$ passos de Monte Carlo, em seguida $f = 0,001$ e $\tau = 1000$ e, por fim, utilizamos o domínio obtido por $f = 0,0001$ e $\tau = 10000$.

Podemos verificar na figura (4.1) que a medida que aumentamos os passos de MC configurações que antes não haviam sido visitadas entram no domínio e poderão ser amostradas e, por conseguinte, terem suas densidades de estados calculadas. Verificamos também que a amostragem nas energias mais baixas (limite superior do gráfico) é mais problemática.

O espaço que contém as configurações acessíveis é peculiar a cada massa de dados. O número de configurações acessíveis cresce com o tamanho da massa de dados e por isso pode-se verificar que a área do domínio é maior nos dados com mais observações, conforme ilustra a figura (4.2).

Apesar de verificarmos uma forma parecida, os limites dependem unicamente dos dados que estão sendo estudados. Em geral, a forma do domínio assemelha-se a um triângulo com um dos catetos praticamente paralelo ao eixo das abcissas, outro cateto inicia com uma curva crescente e segue paralelo ao eixo das coordenadas. O limite inferior, que seria comparável a uma hipotenusa, é uma curva crescente que vai do eixo das coordenadas até o cateto paralelo ao eixo das coordenadas.

A princípio, a região definida pelas magnetizações mínima e máxima e pelas energias mínima e máxima seria retangular, porém o sistema não apresenta energias altas associadas a

magnetizações também altas, pois há uma incompatibilidade uma vez que energias altas são próprias de grandes desordens, ou seja, magnetizações baixas, exceto quando a temperatura é elevada. Por isso, verificamos que a região inferior fica fora do domínio. Desta forma fica definida a região onde deveremos proceder à técnica Wang-Landau para calcular a densidade de estados.

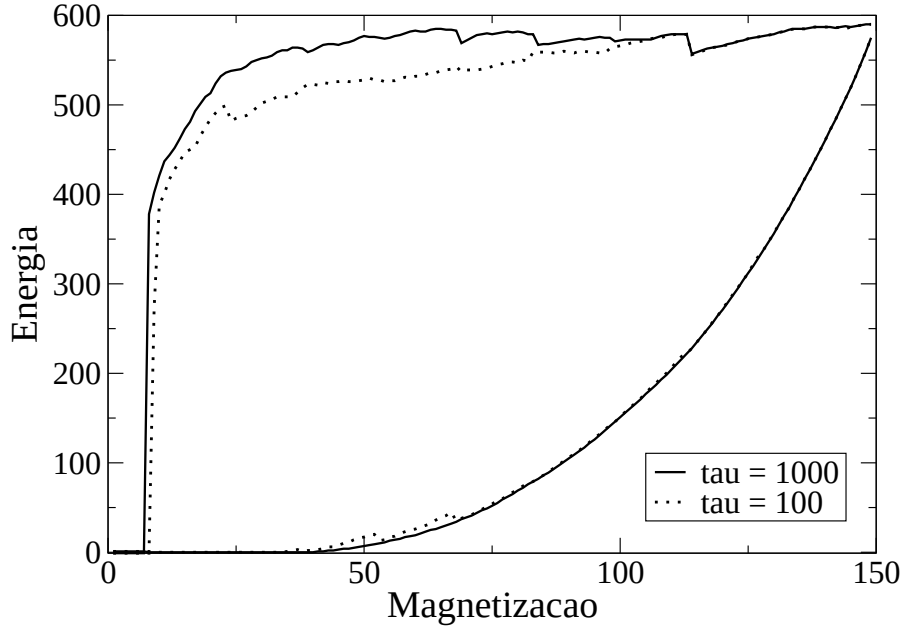

Figura 4.1 Limites da energia e magnetização para a base de dados Íris utilizando $\tau = 100$ e $\tau = 1000$ passos de Monte Carlo e critério de parada $f = 0,001$.

Após obtido o domínio da energia e magnetização para cada sistema físico associado às massas de dados, procedemos ao método Wang-Landau para obter as densidades de estado. De posse da densidade de estados podemos avaliar a energia e o calor específico em função da temperatura e, então, identificar as temperaturas em que a correlação spin-spin deve ser avaliada para finalmente agrupar os dados.

Após identificadas as temperaturas em que existia o interesse em classificar os dados, calculamos a função correlação spin-spin através do algoritmo Metropolis. O número de passos de Monte Carlo (MCS) foi escolhido de acordo com a massa de dados e fomos aumentando os valores até que a classificação ficasse estável, ou seja, obtivéssemos a mesma classificação para uma dada temperatura em diferentes simulações.

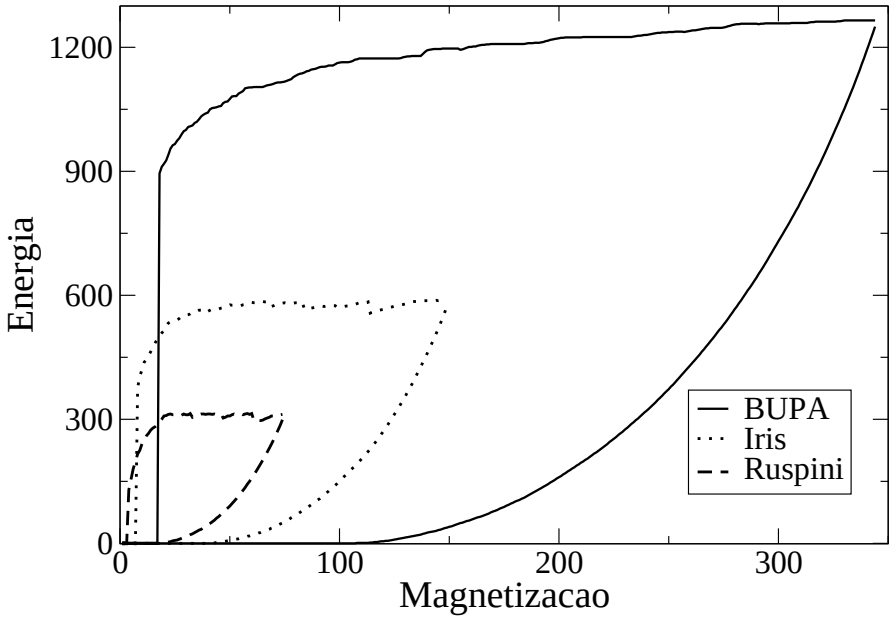

Figura 4.2 Limites da energia e magnetização para as bases de dados BUPA, Iris e Ruspini utilizando $\tau = 10000$ passos de Monte Carlo e critério de parada $f = 0,0001$.

4.1 Aplicações

4.1.1 BUPA

Os dados conhecidos como BUPA Liver Disorders obtidos pela BUPA Medical Research Company constam de uma amostra proveniente de 345 homens solteiros [21]. Os dados consistem de 6 atributos e 2 classes com 200 e 145 indivíduos classificados de acordo com afecções hepáticas. Cinco dos atributos são resultados de um exame de sangue (volume corporal médio, fosfatase alcalina, transferase amino aspartate, aminotransferase aspartata e gama glutamil transpeptidase) e o último atributo é a quantidade de bebida alcoólica ingerida por dia em medida de doses de meio litro.

A figura (4.3) apresenta o comportamento da energia como função da temperatura. Verificamos que para temperaturas acima de cerca de 0,14 o sistema torna-se insensível à temperatura e a energia permanece constante. A partir desta temperatura o sistema degenera e os grupos derretem-se, os objetos tornam-se descorrelacionados.

A figura (4.4) apresenta o comportamento do calor específico como uma função da temperatura para a base de dados BUPA. Cada valor máximo no calor específico está associado a uma "transição de fases". Nas correspondentes temperaturas ocorrem as quebras na estrutura dos dados, tais quebras podem significar que novos grupos estão se formando ou apenas que objetos estão migrando de algum grupo já existente.

A última transição apresenta o maior pico no qual o sistema degenera-se e cada objeto dos dados pertence a um grupo distinto. Os spins correspondentes aos objetos encontram-se completamente decorrelacionados para temperaturas acima de 0,148.

Verificamos ao menos 6 transições de fases, o que nos dá uma ideia de que existem diversos grupos ou existe uma certa hierarquia nos dados. Podemos confirmar a estrutura dos dados procedendo à classificação em faixas distintas de temperatura entre as transições de fases.

Para efetuarmos uma classificação dos dados com o maior número possível de grupos devemos verificar a correlação spin-spin numa faixa de temperatura entre a penúltima e a última transições.

A tabela (4.1) apresenta os agrupamentos obtidos para várias temperaturas antes das transições de fases. Antes da primeira transição de fases, a uma temperatura próxima de 0,04 os dados se dividem em três grupos com 338, 3 e 4 elementos. A medida que a temperatura vai aumentando percebemos que os dados continuam apresentando um grupo grande e alguns elementos deste grupo vão se soltando e formando novos grupos. Após a última transição de fases, a uma temperatura de 0,15, os dados encontram-se muito descorrelacionados e divididos em 49 grupos. Podemos observar que a classificação fornecida pelos pesquisadores não concorda com a aplicação do presente método.

Tabela 4.1 Resultado dos agrupamentos obtidos para a base de dados BUPA em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 15$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^4$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^6$ MCS, cálculo das correlações realizado com $\tau = 10^6$ MCS.

Temperaturas	Agrupamentos
0,04	**3 grupos**: 338;3;4
0,058	**4 grupos**: 332;3;6;4
0,075	**8 grupos**: 315;3;6;2;3;6;7;3
0,1	**11 grupos**: 302;3;3;6;2;4;8;6;5;2;4
0,128	**17 grupos**: 261;16;6;3;3;3;6;8;3;2;5;8;3;7;5;3;3
0,15	**49 grupos**: 4;10;10;33;37;18;5;15;8;2;3;3;2;5;4; 6;4;4;3;2;14;6;2;3;5;4;31;3;6;8;5;5;9;2;4;4;5;11; 4;6;5;6;3;3;3;3;3;2;2

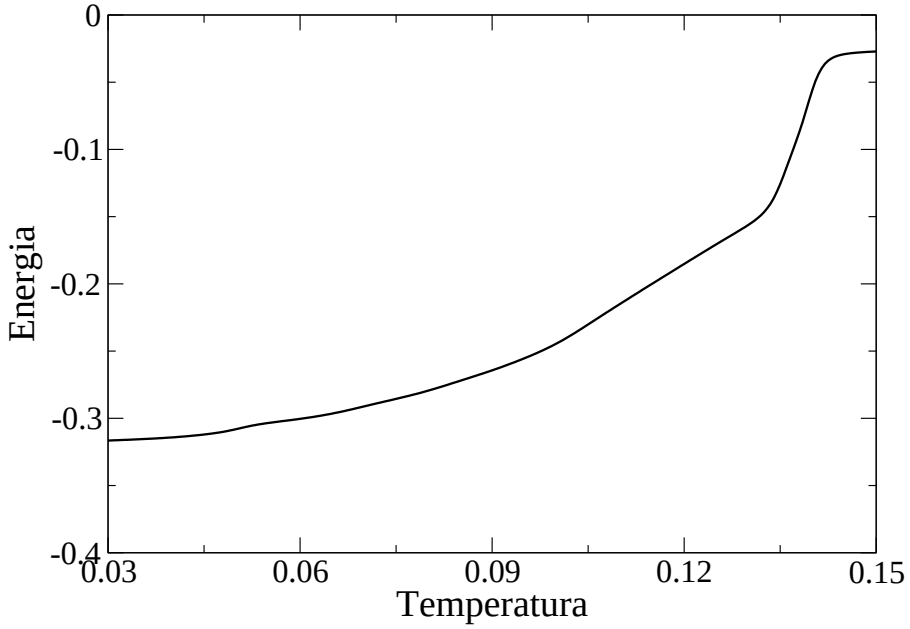

Figura 4.3 Energia em função da temperatura para a base de dados BUPA.

4.1.2 Iris

O conjunto de dados reais da planta Íris, fornecido pela Universidade da Califórnia em Irvine [24], consiste de 150 objetos caracterizados por quatro medidas numéricas que descrevem respectivamente o comprimento da sépala (cm), a largura da sépala (cm), o comprimento da pétala (cm) e a largura da pétala (cm). Esse conjunto é composto de três grupos cada um com 50 objetos. Cada grupo representa um tipo diferente de planta Íris que são chamadas de Íris Setosa, Íris Versicolor e Íris Virginica. O grupo da Íris Setosa é linearmente separado dos grupos das outras duas. Porém os grupos das Íris Versicolor e Íris Virginica não são separados linearmente.

A figura (4.5) apresenta o comportamento da energia como função da temperatura. Verificamos que a partir de uma temperatura próxima de 0,06 o sistema se torna insensível à temperatura e a energia permanece constante. Como o calor específico é a derivada da energia a partir desta temperatura não observaremos nenhum pico na curva de calor específico em função da temperatura.

O comportamento do calor específico em função da temperatura para a base de dados Iris

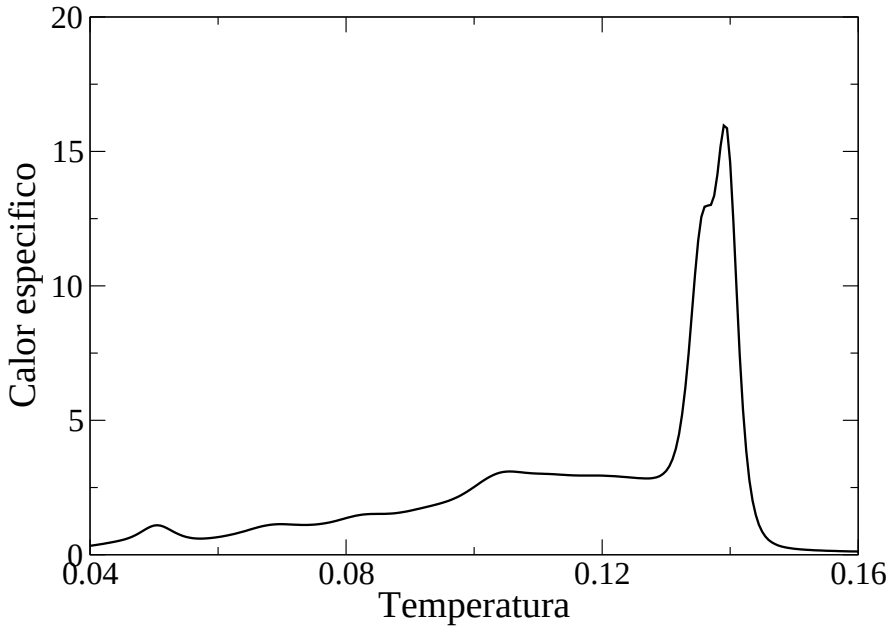

Figura 4.4 Calor específico em função da temperatura para a base de dados BUPA.

é apresentado na figura (4.6). Verificamos dois valores máximos relacionados a transições de fases. A intensidade do pico apresenta relação com o número de ligações rompidas, ou seja o número de spins envolvidos na mudança da estrutura do sistema e, por conseguinte, da estrutura dos dados.

Antes da primeira transição de fases, a uma temperatura de 0,013, observamos a presença de dois grãos com 50 e 100 elementos. Próximo a primeira transição de fases, a uma temperatura de 0,017, dois elementos migram criando um novo grão e observamos, portanto três grupos com 50, 98 e 2 elementos. Tais elementos continuam correlacionados da mesma forma até a temperatura de 0,07 quando verificamos 4 grãos com 50, 94, 4 e 2 elementos. Acima da temperatura em que ocorre a última transição de fases, aproximadamente 0,11, os objetos estão distribuídos em muitos grupos diferentes uma vez que os spins estão muito descorrelacionados. Os agrupamentos obtidos para várias temperaturas são apresentados na tabela (4.2).

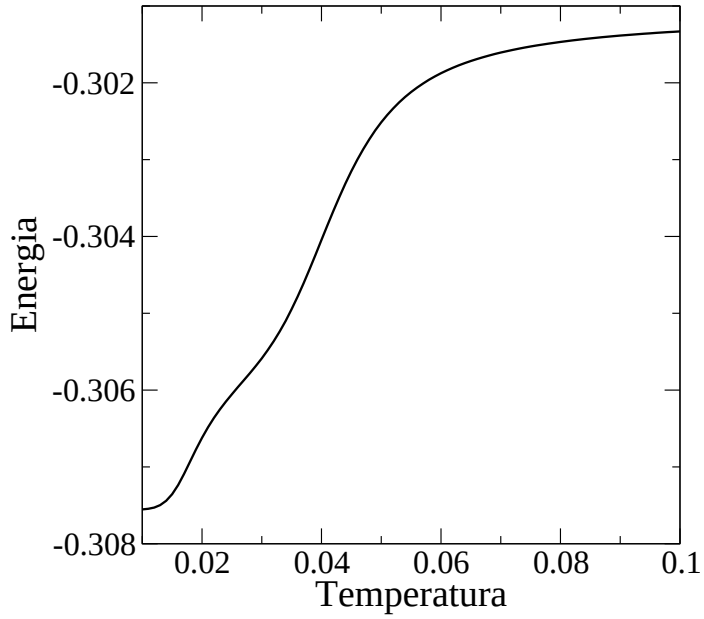

Figura 4.5 Energia em função da temperatura para a base de dados Iris.

4.1.3 Ruspini

A base de dados conhecida como Ruspini [25] foi gerada em 1970 para testes de agrupamentos e é composta de 75 objetos bidimensionais, ou seja, cada objeto é composto por duas variáveis. O conjunto de dados Ruspini é formado por quatro grupos com 20, 23, 17 e 15 elementos em cada grupo.

A figura (4.8) apresenta o comportamento da energia como função da temperatura. Verificamos que até uma temperatura próxima de 0,14 o sistema é insensível à temperatura e a energia permanece constante.

Na figura (4.9) apresentamos o gráfico do calor específico como função da temperatura da base de dados Ruspini. Verificamos dois valores máximos do calor específico que devem estar relacionados a transições de fases. Para confirmar se há uma quebra na estrutura dos dados investigamos a função correlação spin-spin antes de tais picos.

A tabela (4.3) apresenta os agrupamentos obtidos para várias temperaturas. Antes da primeira transição de fases a uma temperatura de 0,077 verificamos através da correlação spin-spin a presença de quatro grãos com 20, 23, 17 e 15 elementos. Após a primeira transição de fa-

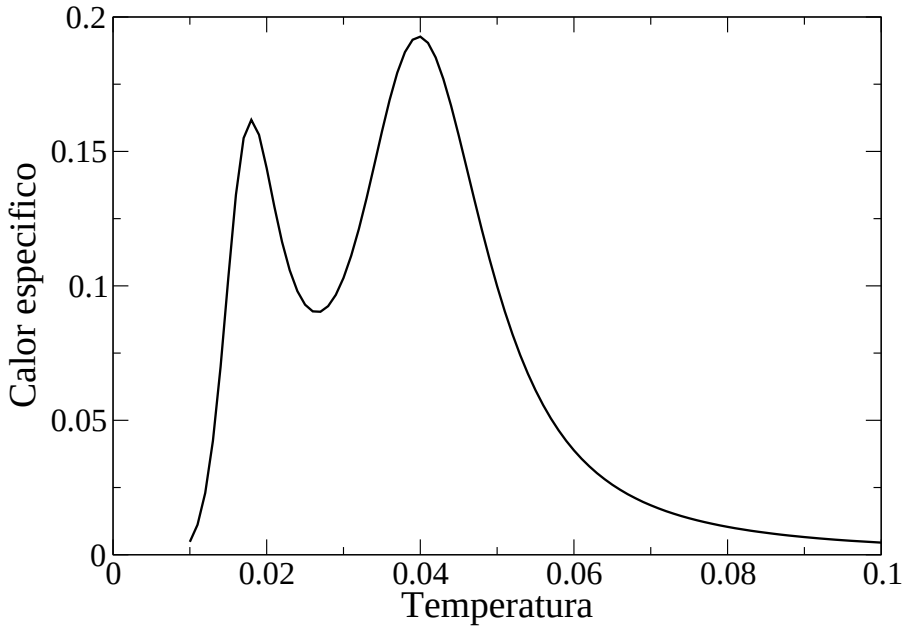

Figura 4.6 Calor específico em função da temperatura para a base de dados Iris.

ses, a uma temperatura de 0,103, surge um novo grão com três elementos formando um novo grupo, e desta forma verificamos 5 grãos com 20, 23, 14, 3 e 15 elementos. A medida que a temperatura aumenta verificamos que os grãos vão se desmanchando, após a segunda transição de fases, a uma temperatura de 0,145, verificamos a presença de 14 grupos. Com o aumento da temperatura verificamos que os objetos estão completamente descorrelacionados, ou seja os objetos estão distribuídos em diversos grupos.

Através da aplicação do presente método verificamos que na temperatura de 0,077, antes da primeira transição de fases, os grupos formados são exatamente os mesmos da classificação verdadeira da base de dados Ruspini. Verificamos também que os três elementos que migram na temperatura de 0,103, próximo à primeira transição de fases, são elementos que apesar de pertencerem a um grupo de 17 elementos estão localizados um pouco distantes dos outros elementos como mostra a figura (4.7).

4.1.4 Cena tridimensional com $\hat{K} = 15$

Tabela 4.2 Resultado dos agrupamentos obtidos para a base de dados Íris em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 15$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^4$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^6$ MCS, cálculo das correlações realizado com $\tau = 10^11$ MCS.

Temperaturas	Agrupamentos
0,013	**2 grupos**: 50;100
0,017	**3 grupos**: 50;98;2
0,026	**3 grupos**: 50;98;2
0,04	**3 grupos**: 50;98;2
0,05	**3 grupos**: 50;98;2
0,08	**5 grupos**: 50;92;4;2;2
0,11	**13 grupos**: 50;46;4;3;4;10;5;13;4;2;5;2;2

Tabela 4.3 Resultado dos agrupamentos obtidos para a base de dados Ruspini em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 15$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^4$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^6$ MCS, cálculo das correlações realizado com $\tau = 10^9$ MCS.

Temperaturas	Agrupamentos
0,077	**4 grupos**: 20;23;17;15
0,103	**5 grupos**: 20;23;14;3;15
0,119	**8 grupos**: 10;4;6;23;8;3;6;15
0,129	**9 grupos**: 10;4;6;23;8;3;4;2;15
0,145	**14 grupos**: 6;4;3;7;20;3;8;3;4;2;7;3;3;2

Geramos 1000 pontos distribuídos uniformemente em uma esfera de raio 0,2 e 1000 pontos distribuídos uniformemente em um toro com raios 0,1 e 0,5. A distribuição dos pontos pode ser observada na figura (4.10).

A figura (4.11) exibe o comportamento da energia em função da temperatura e podemos verificar que até uma temperatura próxima a 0,06 e a partir de uma temperatura próxima a 0,13 o sistema se torna insensível à temperatura e a energia permanece constante. Portanto, na faixa entre estas temperaturas ocorrem os picos na curva do calor específico como função da temperatura.

O comportamento do calor específico em função da temperatura para a base de dados figura tridimensional é apresentado na figura (4.12). Verificamos vários valores máximos relacionados a transições de fases. A intensidade do pico apresenta relação com o número de ligações rompidas, ou seja o número de spins envolvidos na mudança da estrutura do sistema e, por conseguinte, da estrutura dos dados.

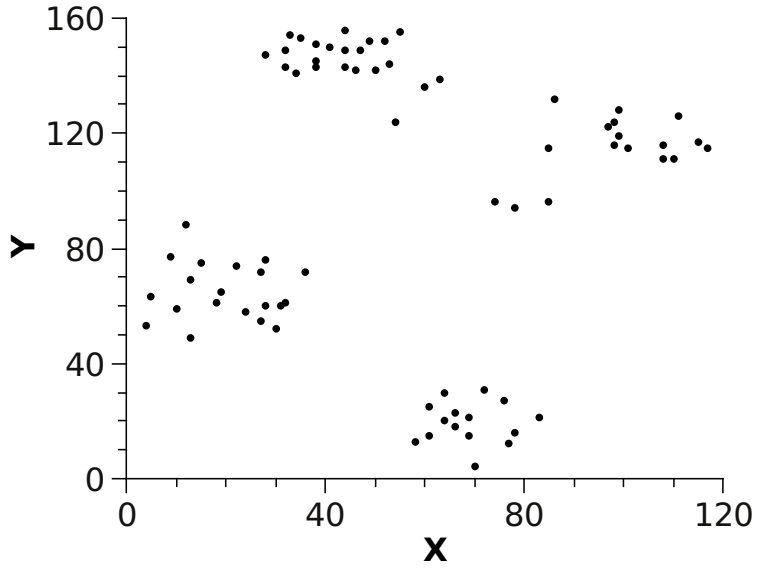

Figura 4.7 Distribuição das observações da massa de dados Ruspini.

Nas temperaturas próximas a 0,065, 0,070, 0,075, 0,078 e 0,081 verificamos picos pouco intensos. O pico mais acentuado é verificado numa temperatura próxima de 0,116. Após esta temperatura verificamos ainda dois valores máximos.

A tabela (4.4) apresenta os agrupamentos obtidos para várias temperaturas inferiores às temperaturas em que ocorrem as transições de fases. Antes da primeira transição de fases, a uma temperatura de 0,05, observamos a presença de dois grãos com 1000 e 1000 elementos. Após esta temperatura um dos grãos vai se dissolvendo. Na temperatura de 0,078 um dos grãos se desfez e o outro grão continua com os 1000 elementos.

4.1.5 Cena tridimensional com $\hat{K} = 10$

Os dados referentes a esta análise são os mesmos gerados para a análise anterior, 1000 pontos pertencentes à uma esfera de 0,2 e 1000 pontos pertencentes a um toro de raios 0,1 e 0,5. O que distingue esta análise é a rede gerada com um número médio de vizinhos mútuos $\hat{K} = 10$.

A figura (4.13) apresenta o comportamento da energia em função da temperatura e podemos verificar que até uma temperatura próxima a 0,03 e a partir de uma temperatura próxima a 0,1 o sistema se torna insensível à temperatura e a energia permanece constante. Portanto, na faixa entre estas temperaturas ocorrem os picos na curva do calor específico como função da

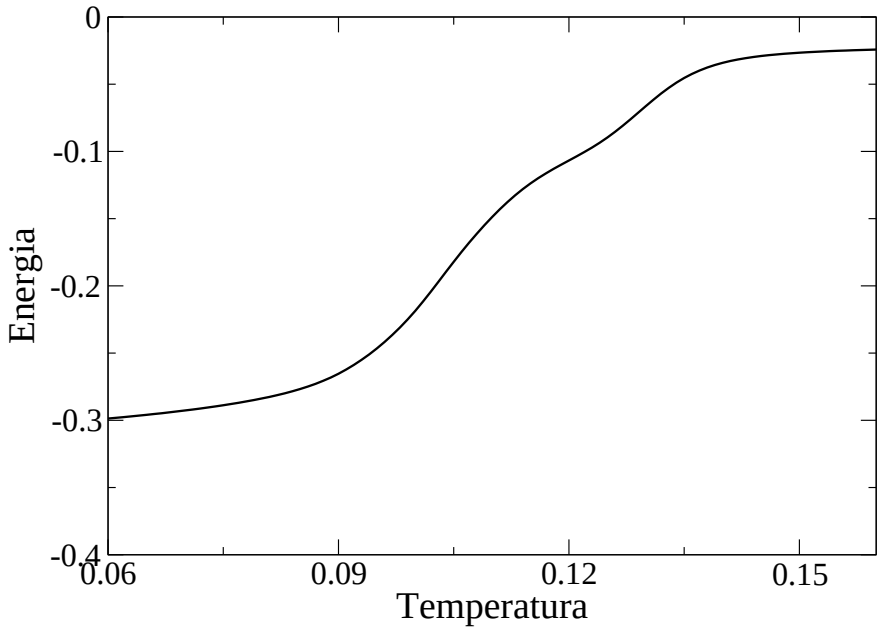

Figura 4.8 Energia em função da temperatura para a base de dados Ruspini.

temperatura.

O comportamento do calor específico em função da temperatura para a base de dados figura tridimensional é apresentado na figura (4.14). Nas temperaturas próximas a 0,02, 0,04, 0,048 verificamos picos pouco intensos, nas temperaturas de 0,06, 0,067, 0,07 e 0,078 verificamos picos mais acentuados e na temperatura de 0,1 verificamos o maior pico no calor específico.

A tabela (4.5) apresenta os agrupamentos obtidos para várias temperaturas inferiores às temperaturas em que ocorrem as transições de fases. Antes da primeira transição de fases, a uma temperatura de 0,026, observamos a presença de dois grãos com 1000 elementos cada. Após a primeira transição de fases, a uma temperatura de 0,032, os dados continuam distribuídos em dois grupos com 1000 elementos cada. Com o aumento da temperatura um dos grupos vai se desfazendo. Na temperatura de 0,075 um dos grãos se desfez e o outro grão continua com os 1000 elementos.

4.1.6 Cena tridimensional com $\hat{K} = 6$

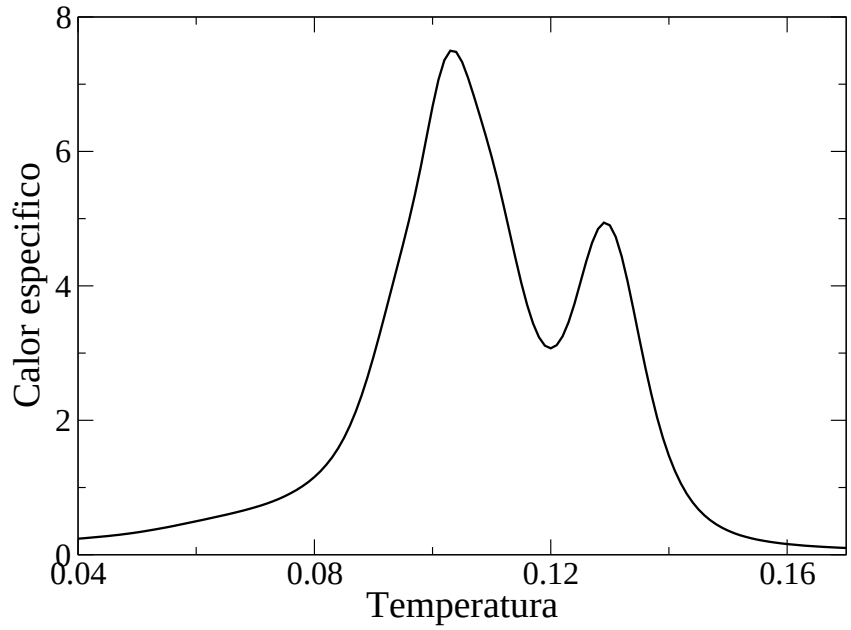

Figura 4.9 Calor específico em função da temperatura para a base de dados Ruspini.

Os dados referentes a esta análise são os mesmos gerados para as duas análises anteriores, 1000 pontos pertencentes à uma esfera de 0,2 e 1000 pontos pertencentes a um toro de raios 0,1 e 0,5. O que distingue esta análise é a rede gerada com um número médio de vizinhos mútuos $\hat{K} = 6$.

A figura (4.15) apresenta o comportamento da energia em função da temperatura e podemos verificar que até uma temperatura próxima a 0,04 e a partir de uma temperatura próxima a 0,16 o sistema se torna insensível à temperatura e a energia permanece constante. Portanto, na faixa entre estas temperaturas ocorrem os picos na curva do calor específico como função da temperatura.

O comportamento do calor específico em função da temperatura para a base de dados figura tridimensional é apresentado na figura (4.16). O gráfico mostra cerca de 23 picos no calor específico associados à transições de fases. Na temperatura próxima a 0,087 verificamos o maior pico no calor específico.

A tabela (4.6) apresenta os agrupamentos obtidos para várias temperaturas. Antes da primeira transição de fases, a uma temperatura de 0,01, observamos a presença de três grãos com 1000, 993 e 7 elementos. Após a primeira transição de fases, a uma temperatura de 0,019, os dados se distribuem em 5 grupos. Com o aumento da temperatura os dois grupos maiores vão

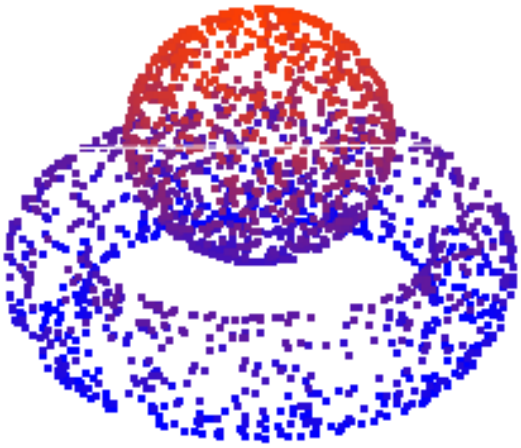

Figura 4.10 Distribuição dos dados objeto tridimensional

se desfazendo. Na temperatura de 0,06 um dos grãos se desfez completamente e o maior grão apresenta 982 elementos.

Percebemos que, apesar dos sistemas obtidos a partir de um número distinto de vizinhos mútuos serem diferentes, o comportamento dos grãos é similar e antes da primeira transição de fases obtivemos a classificação correta, ou no último caso $\hat{K} = 6$, muito próxima da classificação correta. Com o aumento da temperatura, um dos grupos permanece fortemente correlacionado, enquanto o outro grupo se desfaz.

4.1.7 Cena tridimensional com ruído

Os dados referentes a esta análise é composto dos mesmos gerados para as duas análises anteriores, 1000 pontos pertencentes à uma esfera de 0,2 e 1000 pontos pertencentes a um toro de raios 0,1 e 0,5 e acrescentamos um ruído de 1000 pontos distribuídos aleatoriamente no volume que a figura ocupa. Utilizamos para gerar a rede um número médio de vizinhos mútuos $\hat{K} = 8$.

A figura (4.17) apresenta o comportamento da energia em função da temperatura e podemos verificar que até uma temperatura próxima a 0,05 e a partir de uma temperatura próxima a 0,16 o sistema se torna insensível à temperatura e a energia permanece constante. Portanto, na faixa entre estas temperaturas ocorrem os picos na curva do calor específico como função da temperatura.

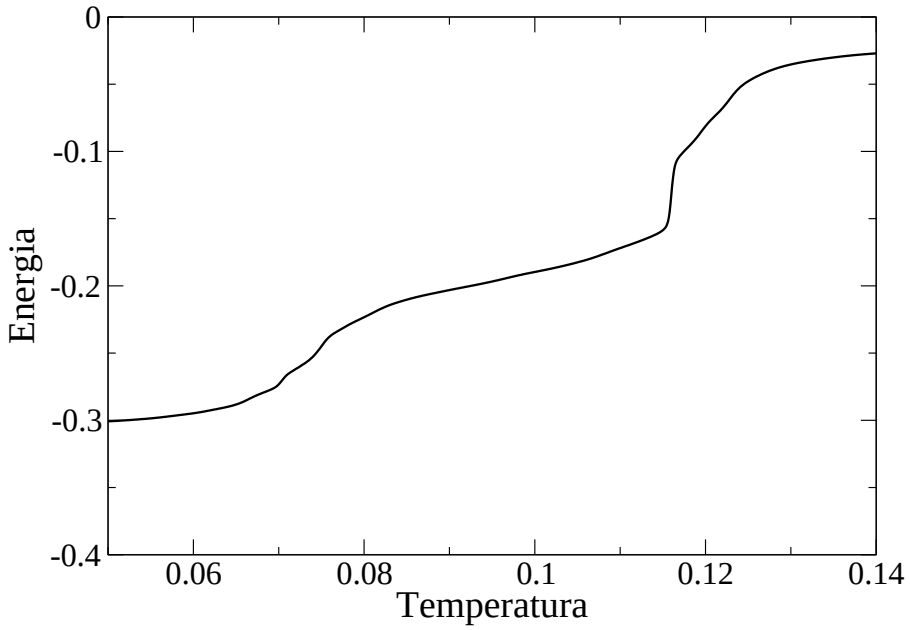

Figura 4.11 Energia em função da temperatura para a base de dados figura tridimensional.

O comportamento do calor específico em função da temperatura para a base de dados figura tridimensional é apresentado na figura (4.18). O gráfico mostra cerca de 22 picos no calor específico associados à transições de fases. Na temperatura próxima a 0,103 verificamos o maior pico no calor específico.

A tabela (4.7) apresenta os agrupamentos obtidos para várias temperaturas. Antes da primeira transição de fases, a uma temperatura de 0,014, observamos a presença de 14 grãos com um grupo maior formado por 2958 elementos. Com o aumento da temperatura o grupo maior vai se desfazendo. Na temperatura de 0,042 percebemos que o grão maior se divide em dois grandes grãos com 1372 e 1121 elementos e os outros elementos se distribuem em 109 grupos menores. A uma temperatura de 0,072 os dados continuam distribuídos em dois grandes grupos com 1305 e 1114 elementos e os outros elementos distribuídos em 164 pequenos grupos.

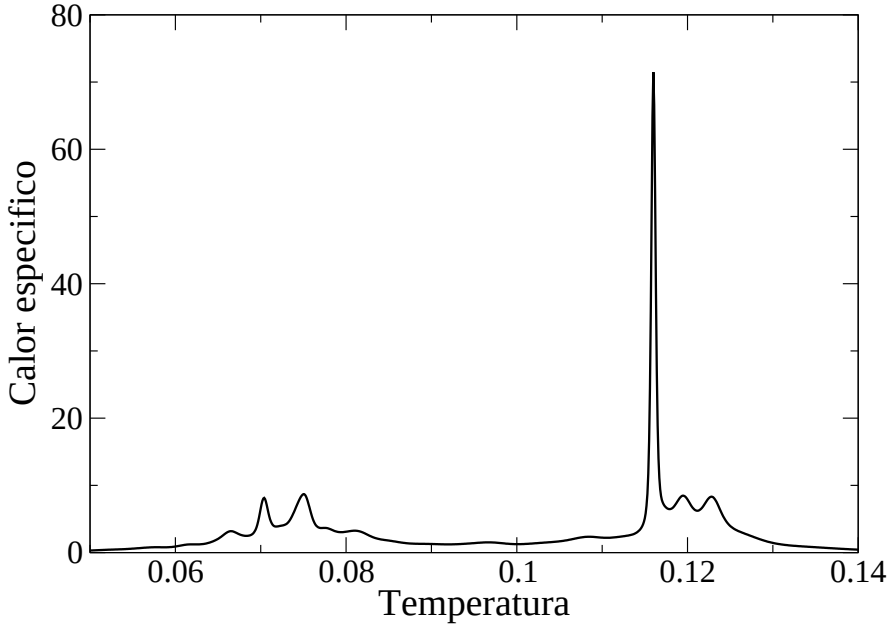

Figura 4.12 Calor específico em função da temperatura para a base de dados figura tridimensional.

Tabela 4.4 Resultado dos agrupamentos obtidos para a base de dados figura tridimensional em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 15$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, cálculo das correlações realizado com $\tau = 10^6$ MCS.

Temperaturas	Agrupamentos
0,05	**2 grupos**: 1000;1000
0,06	**3 grupos**: 998;2;1000
0,068	**18 grupos**: 912;17;19;4;8;9;2;2;3;3;3;2;4;2;3;5;2;1000
0,072	**28 grupos**: 886;17;15;4;4;2;4;7;2;2;2;7;3;3;3;4;2;6;2;2; 2;4;3;6;2;3;3;1000
0,078	**112 grupos**: 101;11;22;38;11;14;2;6;31;4;5;2;24;45;3;3; 13;13;2;8;18;26;12;2;2;3;10;4;51;16;18;23;28;7;4;5;17; 20;40;16;6;2;10;6;7;4;2;5;6;20;7;2;16;2;26;10;7;3;3;5;3; 4;2;8;4;2;10;2;6;2;2;7;5;4;6;2;2;2;6;4;8;3;2;2;3;3;2;2;3; 3;3;2;3;2;3;3;2;2;5;4;3;2;2;3;2;3;3;2;4;3;2;2;3;2;2;1000

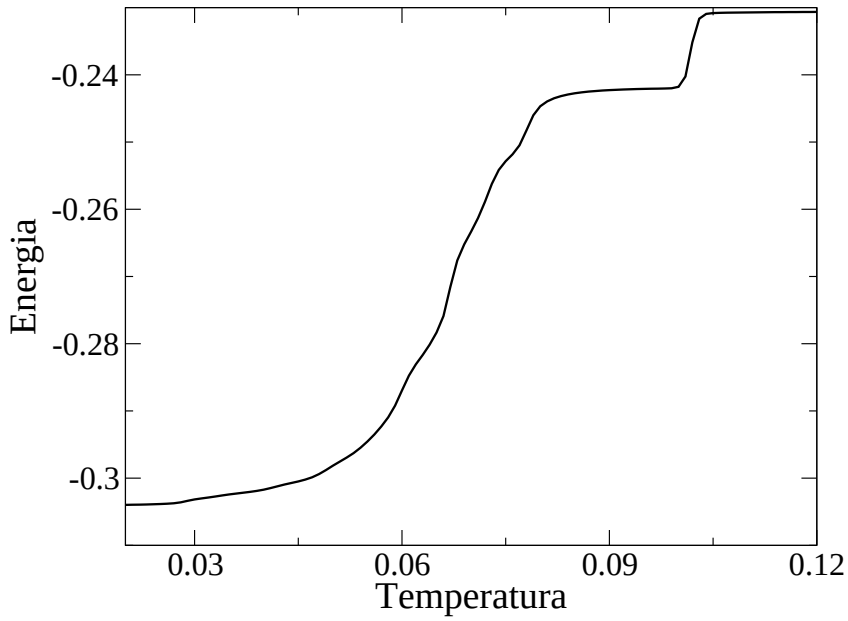

Figura 4.13 Energia em função da temperatura para a base de dados figura tridimensional com $\hat{K} = 10$.

Tabela 4.5 Resultado dos agrupamentos obtidos para a base de dados figura tridimensional em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 10$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, cálculo das correlações realizado com $\tau = 10^6$ MCS.

Temperaturas	Agrupamentos
0,026	**2 grupos**: 1000;1000
0,032	**2 grupos**: 1000;1000
0,044	**4 grupos**: 994;4;2;1000
0,051	**7 grupos**: 987;2;3;2;3;3;1000
0,063	**15 grupos**: 925;17;24;4;8;2;2;3;2;4;2;3;2;2;1000
0,07	**24 grupos**: 904;21;7;2;9;9;2;2;3;2;2;2;4;3;4;2;5; 2;3;3;4;3;2;1000
0,075	**122 grupos**: 94;55;38;31;4;14;10;12;3;5;5;2;2;7; 3;2;3;4;2;9;15;17;18;13;2;2;6;10;4;4;26;2;18;3;4; 18;7;22;4;11;11;13;21;33;17;5;4;3;16;2;10;6;29; 4;13;3;6;2;20;8;2;2;23;15;7;3;6;3;5;16;5;2;6;2;2; 2;10;2;7;2;5;2;7;4;8;9;2;3;2;2;2;3;14;3;3;2;2;4;3; 2;5;2;2;3;2;2;2;2;3;3;4;6;2;4;6;2;2;3;2;2;1000

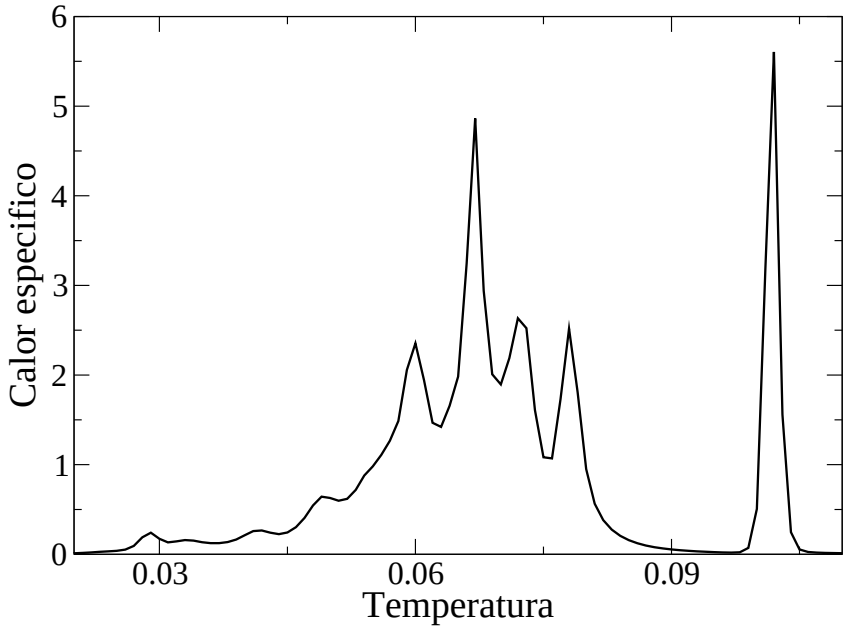

Figura 4.14 Calor específico em função da temperatura para a base de dados figura tridimensional com $\hat{K} = 10$.

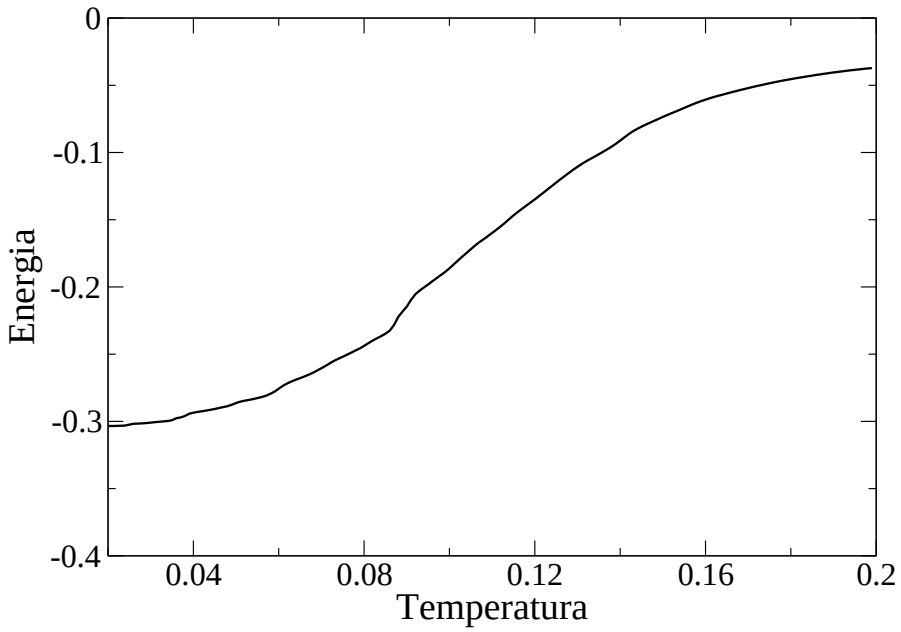

Figura 4.15 Energia em função da temperatura para a base de dados figura tridimensional com $\hat{K} = 6$.

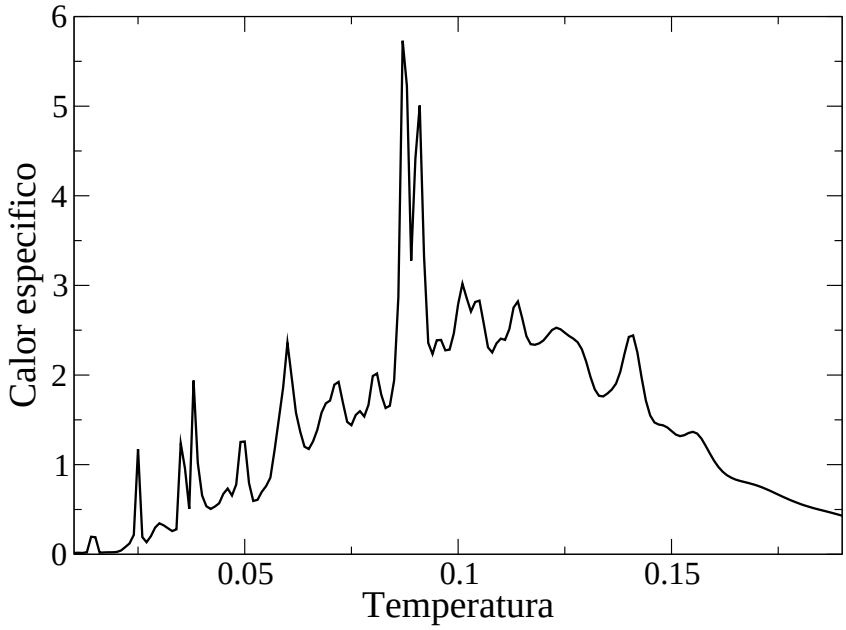

Figura 4.16 Calor específico em função da temperatura para a base de dados figura tridimensional com $\hat{K} = 6$.

Tabela 4.6 Resultado dos agrupamentos obtidos para a base de dados figura tridimensional em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 6$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, cálculo das correlações realizado com $\tau = 10^6$ MCS.

Temperaturas	Agrupamentos
0,01	**3 grupos**: 1000;993;7
0,019	**5 grupos**: 993;4;3;993;7
0,027	**7 grupos**: 987;4;3;2;4;993;7
0,033	**12 grupos**: 975;3;4;3;2;2;4;2;3;2;993;7
0,037	**16 grupos**: 956;3;3;8;4;3;2;3;2;4;2;3;5;2;993;7
0,041	**18 grupos**: 957;3;3;4;3;2;3;2;2;2;4;2;3;5;3;2;993;7
0,047	**33 grupos**: 724;9;53;3;7;2;4;123;6;13;3;3;4;2;3;2;2; 2;3;2;2;2;4;2;3;3;2;5;2;3;2;993;7
0,052	**50 grupos**: 531;98;3;53;3;7;2;2;4;108;9;38;11;4;9;6; 13;3;3;5;4;2;3;6;10;2;6;2;2;3;2;3;2;2;3;2;2;4;3;3;2; 3;3;2;5;2;3;2;993;7
0,06	**103 grupos**: 165;33;44;3;8;9;3;7;2;2;68;3;8;4;58;9;14; 42;2;24;11;6;46;19;6;12;15;4;10;20;31;16;5;8;6;13; 9;6;26;7;7;10;3;3;3;5;2;4;2;7;3;4;3;6;2;6;2;2;2;14;2; 2;2;2;4;5;7;3;2;3;2;2;9;3;3;2;2;4;3;3;6;2;2;2;3;3;3;3; 3;4;2;2;5;7;2;3;3;2;2;982;7;4;7

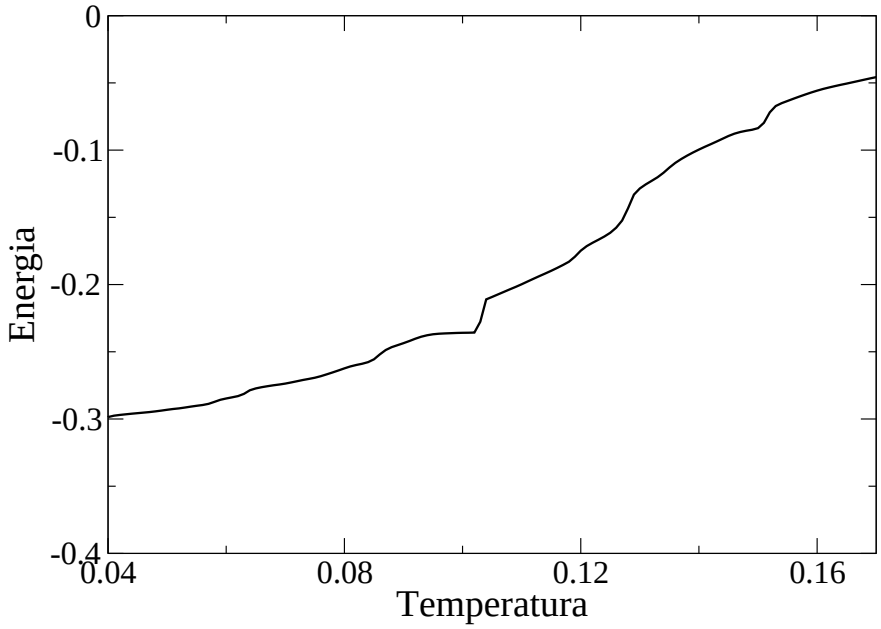

Figura 4.17 Energia em função da temperatura para a base de dados figura tridimensional com ruído.

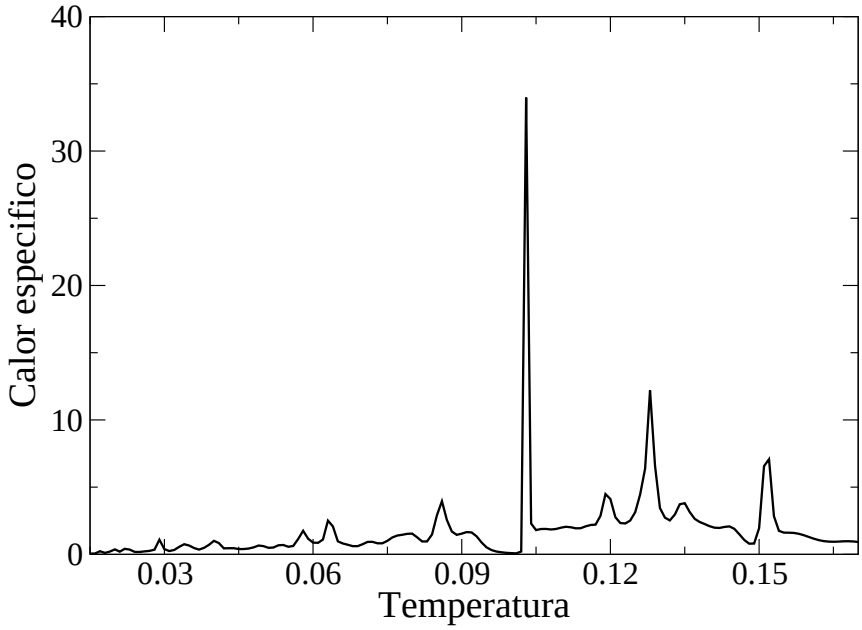

Figura 4.18 Calor específico em função da temperatura para a base de dados figura tridimensional com ruído.

Tabela 4.7 Resultado dos agrupamentos obtidos para a base de dados figura tridimensional com ruído em diversas temperaturas. Rede gerada com número médio de vizinhos mútuos $\hat{K} = 8$, prospecção do domínio realizada com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, Wang-Landau realizado com critério de parada $f = 10^{-4}$ e $\tau = 10^3$ MCS, cálculo das correlações realizado com $\tau = 10^6$ MCS.

Temperaturas	Agrupamentos
0,014	**14 grupos**: 2958;4;9;3;2;5;3;2;2;2;3;3;2;2
0,025	**36 grupos**: 2854;16;3;4;6;11;4;9;3;2;3;3;2;4;7;3;2;5; 15;3;2;2;2;4;2;3;3;2;2;2;3;4;4;2;2;2
0,031	**63 grupos**: 2715;2;7;4;8;4;14;23;4;8;6;6;6;4;3;9;9;10; 3;2;7;3;3;3;2;4;7;3;2;6;10;3;2;3;11;3;2;2;6;9;2;2;3; 2;4;4;2;3;3;2;2;3;2;2;2;2;5;3;4;4;2;2;2
0,042	**111 grupos**: 1372;1121;2;3;5;4;5;9;4;3;16;9;7;4;3;6;4; 10;6;7;6;4;3;9;9;5;3;2;7;6;2;18;14;7;28;3;3;3;2;4;5; 12;2;16;4;5;3;8;2;6;7;6;6;4;3;3;5;3;6;3;7;2;2;2;3;2; 4;2;2;2;3;3;5;3;3;2;2;5;4;4;2;4;4;5;3;3;2;2;3;2;3;3; 2;3;6;2;4;2;2;2;2;5;2;3;4;4;4;2;2;2;2
0,072	**166 grupos**: 1305;2;2;3;10;2;3;2;2;2;3;1114;2;10;4;2; 4;4;3;10;4;3;15;3;8;4;3;2;4;4;4;9;4;5;6;4;5;7;4;2;3; 2;3;7;6;3;2;5;11;5;4;6;3;3;3;2;3;9;4;2;9;4;3;5;3;4; 2;7;4;2;4;3;4;3;3;3;5;5;3;2;6;3;2;2;2;3;3;2;4;2;2;3; 2;2;5;4;3;3;5;3;3;2;2;2;4;4;6;2;3;4;4;4;3;2;3;5;5;3; 2;3;3;3;2;2;3;2;5;2;2;3;3;2;2;3;2;2;2;2;6;2;4;2;2;2; 2;2;2;3;4;4;4;3;2;2;3;2;3;4;2;2;2;2;3;2;2;2

CAPÍTULO 5

Conclusões

O método de agrupamento de dados superparamagnético classificou corretamente as bases de dados bidimensionais e tridimensionais. O que aconteceu fora do esperado é que ao verificar a função de correlação spin-spin a uma temperatura inferior a temperatura da primeira transição de fases os agrupamentos obtidos foram as verdadeiras. Acreditamos que ao gerar a rede completamente conectada as correlações entre os pontos de dados foram armazenadas e a estrutura presente nos dados identificada. Os nossos resultados corroboram com o esperado a respeito da dependência do método com o valor atribuído para o número de vizinhos mútuos, acima de $k = 10$ o método apresentou resultados muito semelhantes, embora o sistema físico seja diferente e, consequentemente, as grandezas termodinâmicas associadas também sejam diferentes.

Ao classificar a base de dados da planta Íris o método apresentou a separação do grupo, que é linearmente independente dos outros dois grupos, a um temperatura inferior a temperatura da primeira transição de fases. Acima da temperatura em que ocorre a última transição de fases verificamos que o grupo linearmente independente continua perfeitamente classificado, o segundo grupo que deveria conter 50 elementos se apresenta dividido em dois grupos com 46 e 4 elementos e o os outros 50 elementos que deveriam pertencer a um único grupo está dividido em 10 pequenos grupos.

Na base de dados BUPA a classificação em dois grupos de acordo com a presença ou não de afecções hepáticas não foi encontrado em nenhuma das temperaturas em que procedemos à classificação. Não sabemos se o método identifica alguma estrutura presente nos dados que levaria a uma classificação diferente, uma vez que o método não se propõe a explicar a natureza dos agrupamentos encontrados, ou se em altas dimensionalidades o método apresenta alguma falha.

Uma dificuldade encontrada na aplicação do método foi a definição da faixa de temperatura em que deveríamos proceder a classificação dos dados, pois transições de fases de primeira ordem em sistemas finitos geram instabilidades e a classificação pode variar com pequenas alterações na temperatura escolhida.

Os resultados obtidos não nos permitem concluir a eficiência do método de agrupamento superparamagnético de dados proposto por Domany et al. [15]. Um próximo passo na investigação seria calcular as densidades de estado bidimensionais $g(E,M)$. Investigações a respeito da interferência do método de Monte Carlo empregado para resolver o sistema físico também podem ser feitas para validar o método, assim como comparar os resultados obtidos a partir de simulações realizadas em diferentes ensembles.

Referências Bibliográficas

[1] Hair, J. F.; Tatham, R. L.; Anderson, R. E.; Black, W. **Multivariate data analysis**. 5.ed. Prentice-Hall, Inc., 1998.

[2] Mardia, K. V.; Kent, J. T.; Bibby, J. M. **Multivariate analysis**. 6.ed. Academic Press, Inc., 1997.

[3] Dillon, W. R.; Goldstein, M. **Multivariate analysis methods and applications**. John Wiley & Sons, Inc., 1984.

[4] Abonyi, J.; Feil, B. **Cluster analysis for data mining and system identification**. Birkhäuser, 2000.

[5] Bishop, C. M. **Neural networks for pattern recognition**. Oxford University Press, 1995.

[6] Dorigo, M.; Stützle, T. **Ant colony optimization**. The MIT Press, 2004.

[7] Dorigo, M.; Stützle, T. **Ant colony optimization: overview and recent advances**. Technical report, TR/IRIDIA/2009-013, IRIDIA, Université Libre de Bruxelles, Bélgica, 2009.

[8] Rosenfeld, A.; Kak, A. C. **Digital picture processing**. 2.ed. Academic Press, Inc., 1982.

[9] Dorai, C.; Jain, A. K. **Shape spectra based view grouping for free-form objects**. In: International Conference on Image Processing (ICIP-95), p. 240243, 1995.

[10] Rasmussen, E. **Clustering algorithms**. In: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419442, 1992.

[11] Fayyad, U. M. **Data mining and knowledge discovery: making sense out of data**. IEEE Expert 11, 5 (Oct.), 2025, 1996.

[12] Witten, I.H.; Frank, E. **Data mining: practical machine learning tools and techniques with Java implementations**. Morgan Kaufmann Publishers, San Francisco, USA, 2000.

[13] Han, J.; Kamber, M. **Data Mining: concepts and techniques**. Morgan Kaufmann Publishers, San Francisco, 2001.

[14] Falkenauer, E. **Genetic algorithms and grouping problems**. John Willey & Sons Ltd., 1997.

[15] Blatt, M.; Wiseman, S.; Domany, E. Super-paramagnetic clustering of data. **Physical Review Letters 76**, p. 3251-3255, 1996.

[16] Newman, M. E. J.; Barkema, G. T. **Monte Carlo methods in statistical physics**. Oxford University Press, 1999.

[17] Landau, D.P.; Binder, K. **A guide to Monte Carlo methods**, Wiley-Interscience, 1986.

[18] Landau, D. P.; Wang, F. A New Approach to Monte Carlo simulations in statistical physics. **Brazilian Journal of Physics 34**, p. 354-362, 2004.

[19] Yeomans, J. M. **Statistical mechanics of phase transitions**. Clarendon Press, 1992.

[20] Tomé, T.; Oliveira, J. O. **Dinâmica estocástica e irreversibilidade**. Editora da Universidade de São Paulo, 2001.

[21] Liver Disorders Data Set. BUPA Dataset. Disponível em: <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/liver-disorders/>. Acesso em: 28 jun. 2010.

[22] Tröster, A.; Dellago, C. Wang-Landau sampling with self-adaptive range. **Physical Review E 71**. 066705, 2005.

[23] Hertz, J.; Krogh, A.; Palmer, R. **Introduction to the theory of neural computation**. AddisonWesley, 1991.

[24] Fisher, R. A. Iris Dataset. 1936, Disponível em: <http://archive.ics.uci.edu/ml/datasets/Iris>. Acesso em: 28 jun. 2010.

[25] Ruspini, E. H. Numerical methods for fuzzy clustering. **Information Sciences 2**, p. 319-350, 1970.