EVERT ELVIS BATISTA DE ALMEIDA

AGRUPAMENTO DE DADOS SUPERPARAMAGNÉTICO

RECIFE-PE – FEVEREIRO/2009.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

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 – Fevereiro/2009.

Ficha catalográfica

A447a Almeida, Evert Elvis Batista de
Agrupamento de dados superparamagnético / Evert Elvis Batista de Almeida. – 2009.
65 f. : il.

Orientador: Adauto José Ferreira de Souza
Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco. Departamento de Estatística e Informática.
Inclui referências bibliográficas.

CDD 574. 018 2

1. Agrupamento de dados
2. Reconhecimento de padrões
3. Simulação no ensemble microcanônico
I. Souza, Adauto José Ferreira de
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

AGRUPAMENTO DE DADOS SUPERPARAMAGNÉTICO

EVERT ELVIS BATISTA DE ALMEIDA

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 26/02/2009 pela Comissão Examinadora.

Orientador: Adauto José Ferreira de Souza

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

Banca Examinadora:

Prof. Dr. Paulo Roberto de Araújo Campos
Universidade Federal Rural de Pernambuco

Prof(a). Dr(a). Tatijana Stosic
Universidade Federal Rural de Pernambuco

Prof. Dr. Tsang Ing Ren
Universidade Federal de Pernambuco

Dedico este trabalho ao meu avô Eufrásio de Souza Alves (*in memoriam*).

Agradecimentos

Agradeço a Deus por todas as oportunidades das quais fui contemplado para chegar até aqui. Agradeço ao meu orientador pela paciência, nestes dois anos de trabalhos. Agradeço Sheila Maria pelo fundamental apoio. Agradeço a minha mãe pelo carinho ao longo de toda vida. Agradeço a esta Universidade que me acolheu deste os tempos de especialização em ensino de matemática. Agradeço aos professores e funcionários que compõe o PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA pela agilidade, atenção e o pelo apoio financeiro que proporcionou a participação em congressos. E um agradecimento todo especial aos amigos Adilton, Erinaldo, Felipe, Juliana, Lidiane e Vanessa.

“Quando você acha que sabe todas as respostas vem a vida e muda todas as perguntas ”

Bob Marley.

Resumo

Aplicamos um método não supervisionado de agrupamento de dados para identificar padrões em vários conjuntos dados. A técnica baseia-se em um mapeamento do problema em um sistema magnético granular heterogêneo, cujo comportamento é investigado através de métodos Monte Carlo comumente empregado no campo da física estatística. Cada objeto é descrito por um conjunto de atributos de valores numéricos, interpretados como um ponto em um espaço euclidiano de dimensão apropriada. O mapeamento consiste em associar a cada item do conjunto, um ponto no espaço, um spin de Potts. O sistema físico é descrito por um hamiltoniano de Potts de muitos estados, no qual a interação entre os spins decai exponencialmente com a distância entre eles. Itens semelhantes, próximos, interagem fortemente enquanto que aqueles mais distantes entre si interagem apenas fracamente. O magneto atinge um estado superparamagnético para temperaturas suficientemente altas, no qual os spins de alguns grãos permanecem fortemente correlacionados, porém, os grãos estão fracamente ligados entre si. Cada grão corresponde a um grupo. Implementamos o método no ensemble microcanônico, no qual a energia total é conservada e constitui o parâmetro de controle. Nesse caso, a temperatura é calculada ao longo do processo e podemos acessar estados termodinamicamente estáveis, metaestáveis, bem como, instáveis. Trabalhamos com três conjuntos artificiais de dados, em duas e três dimensões, e um conjunto de dados reais com quatro dimensões. O desempenho do método foi satisfatório em todos os casos investigados.

Palavras-chave: Agrupamento de dados, reconhecimento de padrões, simulação no ensemble microcanônico.

Abstract

We applied a non-supervisioned data clustering technique based on a map of the problem into an inhomogeneous granular magnet problem. The physical behavior of the magnet is studied through the usual Monte Carlo method. Each data item is described by a set of numerical attributes, interpreted as points in a multiple-dimensional Euclidian space. The mapping consists in associating a Potts spin to each data point. The physical system is described by a disordered Potts Hamiltonian with several states with an exponentially decaying interaction among spins. The magnet reaches a superparamagnetic state at high temperatures in which the spins in certain grains are strongly correlated whereas the grains are loosely linked. In this way, each grain corresponds to a group or cluster. We implemented the method in a microcanonical ensemble where the conserved total energy is the control parameter. The temperature is calculated during the simulation and, besides thermodynamic stable states, it is possible to sample unstable and metastable state as well. We work with three artificial multiple-dimensional data set and a four-dimensional real data set. We obtained good results in all cases and discuss some issues concerning the microcanonical implementation of the superparamagnetic data clustering.

Keywords: Data clustering, pattern recognition, microcanonical ensemble simulation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em muitas áreas de conhecimento e aplicações, existe a necessidade de armazenar e manipular grandes quantidades de dados. Em uma fase exploratória precisamos realizar a análise dos agrupamentos formados pelos itens armazenados com o objetivo de obter informações úteis a resolução dos mais diversos problemas. Análise de agrupamento ou “clustering” é uma importante ferramenta nas técnicas exploratória de dados. Os métodos de agrupamentos objetivam separar ou organizar os dados em classes. Trata-se de um procedimento matematicamente mal definido, pois não se sabe a priori o número de classes. A necessidade de agrupar dados aparece em uma grande variedade de áreas científicas com mais diversas finalidades, tais como reconhecimento padrão, inteligência artificial, astrofísica e outros. Como exemplo ilustrativo da técnica, imagine o seguinte experimento: uma criança, que nunca tenha antes visto um canguru ou uma girafa é exposta a centenas de imagens destes animais, sem qualquer explicação previa. Depois de ver um número suficientemente grande de girafas e cangurus a criança terá uma forma bastante clara de compreensão a respeito destas duas diferentes de criaturas. Vemos então, que a criança aprendeu algo novo, sem ser instruída. Esta forma de aprendizagem não supervisionada é provavelmente o mais importante meio de adquirir e processar um fluxo incessante de informações, que atinge os nossos sentidos a partir do mundo que nos rodeia. Nosso cérebro pode aprender sem orientação de um professor, por agrupamento através observações semelhantes. O exemplo definiu uma técnica de reconhecimento de padrões desenvolvido pelo ser humano.

Reconhecimento de padrões engloba uma literatura tão vasta que sua definição é polêmica, e está ligada a busca de “regularidades”. Deste os tempos pré-históricos, a humanidade busca “regularidades” nas quais possa confiar, transmitindo-lhe uma sensação de segurança num mundo hostil. Reconhecer rostos, compreender palavras faladas e escritas, identificar objetos utilizando os sentidos biológicos entre outros processos característicos do ser humano podem ser chamados de reconhecimento de padrões. As ações relacionadas com a categorização foram cruciais para nossa sobrevivência ao longo do tempo, e temos

evoluído de maneira impressionante criando sofisticadíssimos sistemas neurais para o desenvolvimento de tarefas complexas.

Reconhecer padrões é uma característica de todos os organismos vivos. Contudo, criaturas podem fazer uso de diferentes formas para reconhecer padrões. O homem pode reconhecer outro homem através da voz, fisionomia, caligrafia entre outros. Por outro lado um cão é capaz de desempenhar o mesmo processo através do cheiro. Nos exemplos citados observamos o uso de duas ou mais estratégias para realizar a mesma função.

A estatística utilizada em reconhecimento de padrões é uma aplicação computacional que faz uso de vários conceitos tais como probabilidade e estimativa entre outros. O uso de técnicas de reconhecimento de padrões se faz necessário em numerosos campos da ciência como medicina, visão computacional, robótica, sistemas militares, finanças entre outros. Algumas destas são citadas, como segue:

- Um médico faz o diagnóstico de uma doença relacionando os sintomas e baseando se em resultados de exames.
- Um radiologista localiza áreas onde existe tecido não sadio em imagens de raios-X.
- Um analista militar classifica regiões de uma imagem para uso em sistemas de segmentação.
- Um geólogo determina se um sinal sísmico representa um iminente terremoto.
- Um gerente de empréstimo de um banco deverá decidir se um cliente apresenta um risco de crédito com base no seu rendimento, histórico de crédito e outras variáveis.
- Um fabricante deve classificar a qualidade dos materiais antes de usar em seus produtos.

São amplas as aplicações de técnicas de reconhecimento de padrões em problemas que para os sentidos humanos aparenta uma fácil solução, mas para ser resolvida por um computador pode apresentar um alto grau de complexidade. Dentre estes problemas podemos citar o reconhecimento da fala ou de caracteres. Como exemplo de reconhecimento de caracteres, podemos considerar uma situação hipotética, como distinguir alguns caracteres capturados em uma imagem. Uma representação conveniente da imagem é uma matriz cujos elementos assumem um dos valores inteiros 0 ou 1 (0 para branco e 1 para preto). Para construir tal

representação, um reticulado é posto sobre a imagem e fazemos cada célula deste corresponder a um elemento da matriz, ver figura 1. As células do reticulado são comumente chamadas de *pixels*, a partir da expressão em inglês, *picture elements*. A representação final de cada caractere é obtida através da soma dos elementos de cada coluna da matriz, que resulta em um histograma como exemplificado na figura 1 para as letras a, b e c respectivamente:

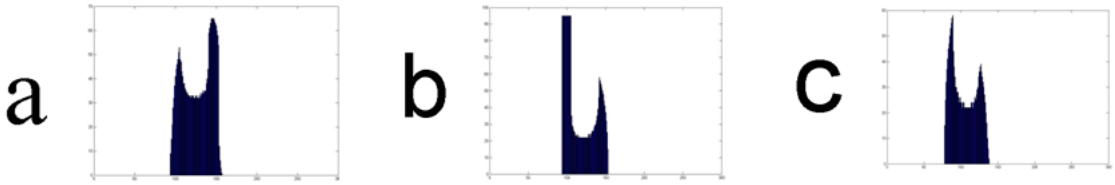


Figura 1.1Histograma da projeção da matriz de pixels correspondentes as letras **a**, **b** e **c**.

Observe que cada letra possui um histograma característico e através dele podemos realizar um processo automatizado para executar o reconhecimento. Mais exemplos da utilização de histogramas para reconhecimento de padrões em caracteres podem ser visto em Duda (2003), Bishop (1995) e Fukunaga (1990). Na realidade em reconhecimento de padrões temos uma vasta coleção de problemas variados cujas soluções empregam técnicas importadas de outras áreas e são projetadas para atender a especificidade do problema particular. Técnicas que tiveram origem na Física foram propostas como soluções para problemas de agrupamentos de dados como podemos ver em Domany (1996) e Marangi (2000). Até o comportamento de insetos serviram para fundamentar métodos de agrupamentos de dados como podemos ver em Dorigo (2004). As contribuições de reconhecimento de padrões atendem os mais diversos campos como teoria de autômato celular, teoria da decisão, teoria da informação, teoria das probabilidades, redes neurais, inteligência artificial dentre outras. Reconhecimento de padrões é entendido como a caracterização de dados de entrada em classes identificáveis através de extração de características ou atributos fundamentais. A especificação de características para uma boa classificação depende do problema específico que se quer resolver, e especificá-las é mais arte que propriamente ciência. Os padrões ou objetos analisados são habitualmente representados por vetores, descrito por uma seqüência de números reais dentro de um espaço multidimensional. Isto é, a n-tupla de valores reais pode ser interpretada como um vetor $\vec{x}$ ou um ponto de $x \in \Re^n$.

Muitas das técnicas exigem uma medida de dissimilaridade entre os objetos ou distâncias entre os vetores para executar a partição dos grupos.

Nesta dissertação iremos aplicar uma técnica de agrupamento não supervisionada há vários conjuntos de dados. O método consiste no mapeamento do problema em um sistema magneto granular heterogêneo, o qual é posteriormente investigado através de uma técnica de Monte Carlo. O mapeamento consiste em associar um spin de Potts a cada item de dado. Os spins interagem aos pares e a intensidade desta interação decai com a distância que os separa. Em altas temperaturas o sistema é completamente desordenado, enquanto que ele apresenta ordem magnética em baixas temperaturas. Para temperaturas intermediárias, o magneto atinge um estado superparamagnético no qual apenas certos aglomerados de spins apresentam ordem magnética. Ou seja, grãos de spins fortemente correlacionados são formados. Cada grão corresponde a um grupo. Realizamos a mecânica estatística do sistema modelo através de um método Monte Carlo no qual a energia é mantida constante, ou seja, em um ensemble microcanônico.

Esta dissertação está organizada como segue. Neste capítulo daremos algumas definições necessárias ao entendimento da técnica aplicada em agrupamento de dados. No segundo capítulo faremos uma exposição a respeito de tópicos de física e métodos computacionais relacionados com a técnica em estudo. No terceiro capítulo apresentaremos a aplicação da técnica em alguns bancos de dados artificiais utilizados anteriormente na implementação de outros métodos de agrupamentos que tiveram origem em técnicas usadas na física. Finalmente no quarto capítulo teceremos alguns comentários sobre os resultados obtidos com a aplicação do método.

1.1 DEFINIÇÕES

1.1.1 DISTÂNCIAS E MEDIDAS DE DISSIMILARIDADE

A distância entre dois pontos é medida pelo comprimento do segmento de reta que os liga. Quando se fala na distância entre dois pontos da superfície da Terra, então a distância é o comprimento do menor caminho entre todos os possíveis, sobre a superfície partindo de um ponto e atingindo o segundo. A distância é sempre uma medida positiva e tem a propriedade de que a distância de um ponto A até um ponto B é idêntica à distância do ponto B até o ponto A. A idéia de distância entre dois pontos é formalizada e generalizada pela matemática através do conceito de métrica. Um espaço onde há uma distância ou métrica definida é chamado de espaço métrico. Mais precisamente, se $\mathbb{S}$ é um conjunto, uma métrica em $\mathbb{S}$ é uma função $\mathbf{d}: \mathbb{S} \times \mathbb{S} \to \mathbb{R}$, chamada de distância (ou dissimilaridade) que associa dois elementos de $\mathbb{S}$ a um número real. Qualquer função $\mathbf{d}$ que satisfaça aos axiomas abaixo, é uma distância:

1- Ser positivamente definida $\mathbf{d}(\mathbf{x}, \mathbf{y}) > 0$ para todos os x, y $\in \mathbb{S}$
2- Ser simétrica $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \mathbf{d}(\mathbf{y}, \mathbf{x})$ para todos os elementos de x, y $\in \mathbb{S}$.
3- Obedecer a desigualdade triangular. Para todos os x, y, z elementos de $\mathbb{S}, \mathbf{d}(\mathbf{x}, \mathbf{y}) \leq \mathbf{d}(\mathbf{x}, \mathbf{z}) + \mathbf{d}(\mathbf{z}, \mathbf{y})$.
4- Ser nula apenas para pontos coincidentes. $\mathbf{d}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \Leftrightarrow \mathbf{x} = \mathbf{y}$

O objetivo destas métricas é medir similaridade (correlação) e dissimilaridade (distância) entre vetores. O fato de se distinguir entre medidas de similaridade e dissimilaridade é apenas uma questão de terminologia, já que os dois conceitos são recíprocos. Contudo, a idéia por trás desta dicotomia é que a correlação aumenta à medida que a similaridade entre vetores aumenta, enquanto que distâncias diminuem. Freqüentemente nos deparamos com o seguinte problema, duas ou mais características escolhidas para representar um objeto ou evento, são influenciadas por um mecanismo que tendem a variar juntas. Esta correlação existente degrada o desempenho do processo de classificação baseado principalmente na distância euclidiana. Outra situação semelhante ocorre quando as características possuem escalas bem distintas. Veremos mais adiante como podemos lidar com este

problema. É possível definir várias funções *d* que satisfazem os axiomas acima e todas serviriam para medir a dissimilaridade entre objetos. A mais usada é a conhecida distância Euclidiana, porém várias outras medidas de distâncias podem e são utilizadas para a identificação de grupos. A escolha de uma determinada função distância deve ser criteriosa, pois os resultados do agrupamento e eficiência do método empregado podem depender da métrica utilizada.

1.1.2 MATRIZ DE DADOS E PADRONIZAÇÃO DOS DADOS

Considere que sejam avaliados p atributos de cada um dos n objetos de um determinado conjunto. É conveniente organizar os dados em uma matriz $n \times p$, na qual cada linha corresponde a um objeto. Isto é, o f-ésimo atributo do i-ésimo objeto é denotado por x_{if} (em que i=1,..., n e f=1,..., p), e a matriz dados é representada da seguinte forma:

$$\begin{pmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{pmatrix}$$

É claro que se uma das variáveis for medida em uma escala muito diferente das outras, ela pode dominar o resultado do cálculo das distâncias ou, ao contrário, não exercer qualquer influência. Por exemplo, para tomar um caso extremo, suponha que n homens estão sendo comparados, e que uma variável avaliada é a estatura e a outra as dimensões dos dentes, se todas as medidas estão em milímetros. Diferenças na estatura estarão na ordem de talvez 20 ou 30 mm, enquanto que diferenças nas dimensões dos dentes estarão na ordem de 1 ou 2mm. Nos cálculos simples fornecerão então distâncias entre indivíduos que são essencialmente diferenças de estatura, com diferenças nos dentes tendo efeitos desprezíveis. Para evitar problemas de escalas diferentes ou que as unidades escolhidas para mensurar as variáveis possam afetar a similaridade entre os objetos, fazendo com que as variáveis contribuam de forma mais igualitária para similaridade dos objetos, utiliza-se um procedimento chamado padronização.

Usualmente as variáveis são padronizadas de algumas maneiras antes das distâncias serem calculadas, de modo que todas as p variáveis sejam igualmente importantes na determinação destas distâncias. Isto pode ser feito re-escalando as

variáveis de modo que a média sejam todas nulas e as variâncias unitárias. Alternativamente, cada variável pode ser re-escalada para ter um valor mínimo igual a zero e valor máximo igual a um. Infelizmente, padronização tem efeito de minimizar diferenças de grupo, pois se os grupos são bem separados pelas variáveis, então a variância desta variável será grande.

Temos várias funções para padronizar as variáveis como podemos ver em Frei (2006), mas vamos citar apenas a que foi utilizada neste trabalho dada por:

$$\boldsymbol{x_{if} = \frac{x_{if} - \overline{x}_f}{s_f}} \qquad (\mathbf{1.1})$$

Onde:

$$\boldsymbol{\overline{x}_f = \frac{1}{n}\sum_{i=1}^{n} x_{if}} \qquad (\mathbf{1.2})$$

e

$$\boldsymbol{s_f = \left(\frac{1}{n}\sum_{i=1}^{n}\left(x_{if} - \overline{x}_f\right)^2\right)^{\frac{1}{2}}} \qquad (\mathbf{1.3})$$

Nas simulações que serão realizadas posteriormente as variáveis dos dados terão aproximadamente a mesma influência no cálculo da distância. Isto foi obtido por escalonamento subtraindo cada variável da média e dividindo pelo desvio padrão para os n indivíduos comparados.

1.1.3 DISTÂNCIA MINKOWSKI

É definida como a distância entre dois elementos $x_i \; e \; x_k$, $\mathrm{k} \neq \mathrm{i}$, e dada pela expressão:

$$d(x_i, x_k) = [\sum_{f=1}^{p}\left(x_{if} - x_{kf}\right)^{\lambda}]^{\frac{1}{\lambda}} \qquad (1.4)$$

Esta distância apresenta uma forma generalizada para tratar outras medidas de distâncias como a Euclidiana quando o $\lambda = 2$, Manhattan para $\lambda = 1$ e Chebyshev para $\lambda = \infty$. A Minkowski $\lambda \geq 3$ é menos afetada pela presença de valores discrepantes na amostra do que a distância Euclidiana. A distância Euclidiana é mais usada nos mais diversos métodos de reconhecimento de padrões como podemos ver em Hammer (1996). Dada por uma hiper-esfera, possui a

propriedade de dar maior ênfase a maior diferença entre uma única variável. Vemos nos exemplos abaixo os contornos apresentados pela distância Minkowski ao variarmos o parâmetro λ.

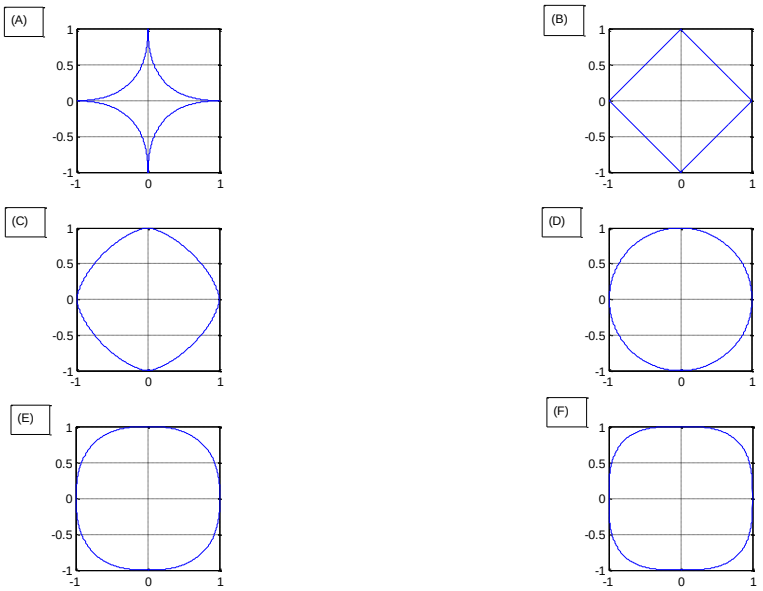


Figura 1.2 Representação dos contornos gerados variando o valor de λ. Figura 1.2.A $\lambda = 0.5$, figura 1.2.B $\lambda = 1$ (distância Manhattan), figura 1.2.C $\lambda = 1.5$, figura 1.2.D $\lambda = 2$ (distância Euclidiana), figura 1.2.E $\lambda = 2.5$ e figura 1.2.F $\lambda = 3$.

Os contornos desenhados nos gráficos são curvas de nível da figura 1.2, onde podemos afirmar que fazendo $\lambda = 1$ ou $\lambda \to \infty$ temos quadrados ou caixas e fazendo $\lambda = 2$ teremos uma esfera. A escolha da melhor distância pode levar a elevados níveis de classificação corretas, contudo em alguns casos temos insucessos no processo sendo assim podemos atribuir razões para este fato como:

- Características inadequadas que não conseguem distinguir as classes.
- Correlações entre as características.
- Existência de subclasses distintas embutidas nos dados.
- Espaço de padrões pode ser complexo demais.

1.1.4 CORRELAÇÕES

Ao estudarmos duas ou mais variáveis simultaneamente, é de interesse estudar se existe ou não dependência ou se uma influência o comportamento da

outra. Existem várias medidas para representar a dependência entre variáveis aleatórias como covariância e coeficiente de correlação. A covariância também costuma ser chamada de correlação entre as variáveis envolvidas, mas alguns autores preferem reservar a palavra correlação para se usa a seguinte normalização:

$$\boldsymbol{corr}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\langle \boldsymbol{X},\boldsymbol{Y} \rangle}{\sqrt{\boldsymbol{var}(\boldsymbol{x})\boldsymbol{var}(\boldsymbol{y})}} = \frac{\langle (\boldsymbol{X} - \langle \boldsymbol{X} \rangle)(\boldsymbol{Y} - \langle \boldsymbol{Y} \rangle) \rangle}{\sqrt{\boldsymbol{var}(\boldsymbol{x})\boldsymbol{var}(\boldsymbol{y})}} \qquad (\mathbf{1.5})$$

Em sistemas interagentes de partículas as simulações numéricas permitem cálculo de grandezas macroscópicas ou termodinâmicas. Uma delas é função de correlação entre as partículas. Essa grandeza é fundamental para o conhecimento do estado de agregação ou ordenamento das partículas que compõem o sistema. Para o sistema de Ising, definimos a função de correlação $\boldsymbol{G}(\boldsymbol{r})$ por

$$\boldsymbol{G}(\boldsymbol{r}) = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{r'}} \langle \boldsymbol{\sigma}_{\boldsymbol{r'}} \boldsymbol{\sigma}_{\boldsymbol{r'}+\boldsymbol{r}} \rangle \qquad (\mathbf{1.6})$$

Dada uma configuração de spins, geradas pela simulação, essa função é obtida da seguinte forma. Para cada par de spins separados por um vetor $\boldsymbol{r}$, somamos +1 se os dois spins forem de mesmo sinal e -1 se for de sinais contrários. Podemos observar nas figuras 1.3, 1.4, 1.5, 1.6 e 1.7 as ilhas de correlações geradas quando variamos a temperatura do sistema, no estado inicial vemos que os spins estão distribuídos aleatoriamente como vemos na figura 1.3, simulando o sistema vemos a formação de pequenas ilhas de correlação entre os spins. Com o aumento da temperatura as ilhas de correlação crescem, e verificamos uma mudança de fase no sistema figuras 1.4, 1.5 e 1.6. Com o aumento sucessivo da temperatura vemos que o sistema muda completamente de fase até que o sistema esteja totalmente descorrelacionado figura 1.7.

A correlação entre variáveis ou partículas assume valores que estão entre -1 e 1. Quando os valores de $|\boldsymbol{corr}(\boldsymbol{X},\boldsymbol{Y})| \approx \mathbf{1}$ demonstram que as variáveis variam simultaneamente, e para $|\boldsymbol{corr}(\boldsymbol{X},\boldsymbol{Y})| \approx \mathbf{0}$ temos uma variação independente não havendo correlação.

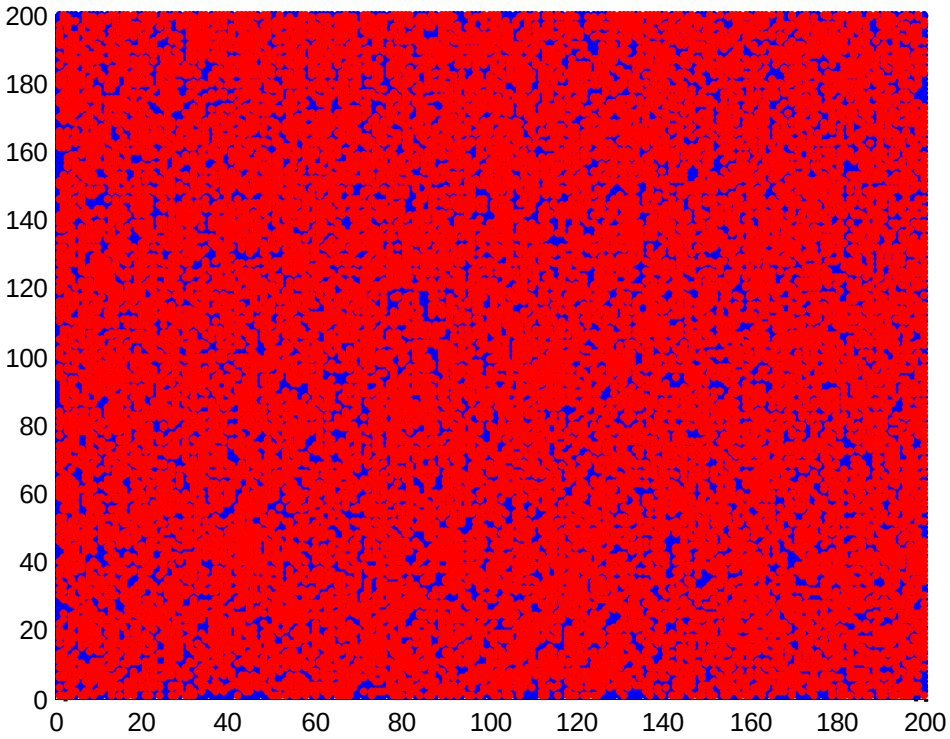


Figura 1.3. Modelo de Ising em sua configuração inicial de uma rede 200X200.

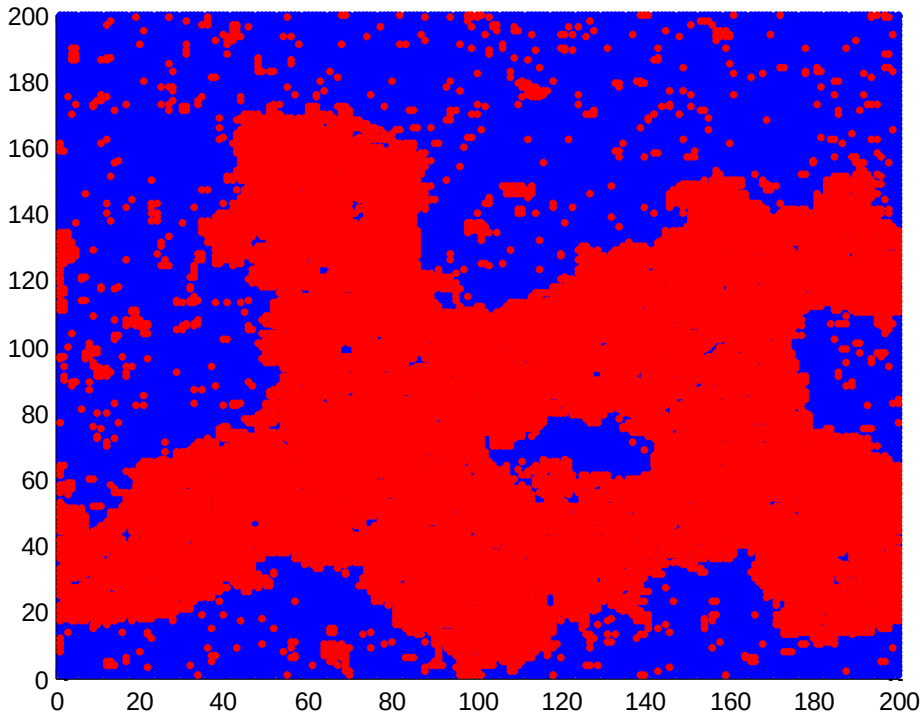


Figura 1.4. Modelo de Ising simulado em uma temperatura T=2, com 1000 passos de Monte Carlo.

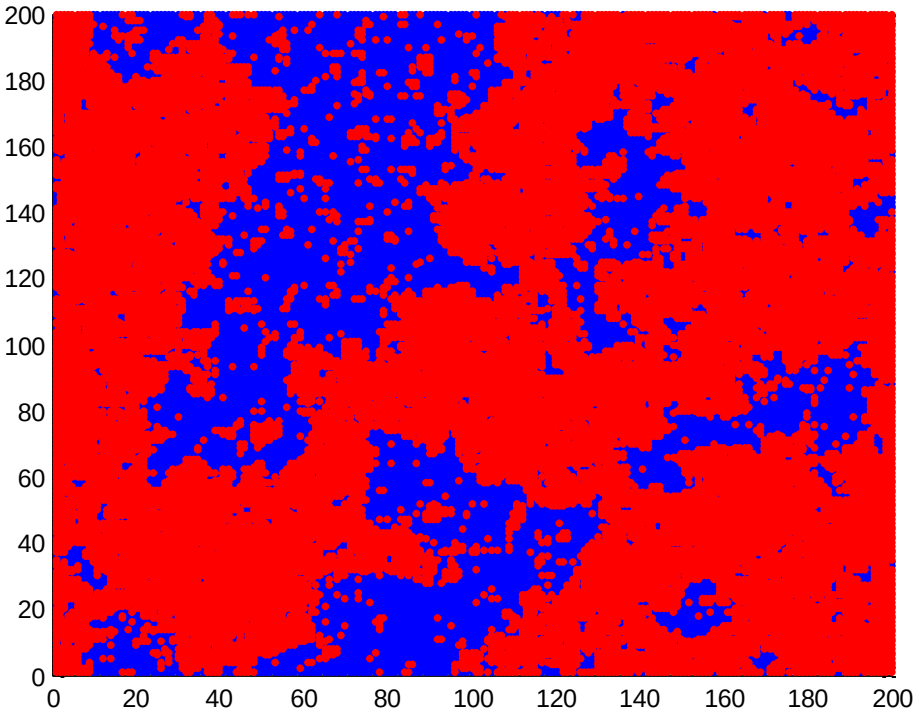


Figura 1.5. Modelo de Ising na temperatura de T=2.4, com 1000 passos de Monte Carlo.

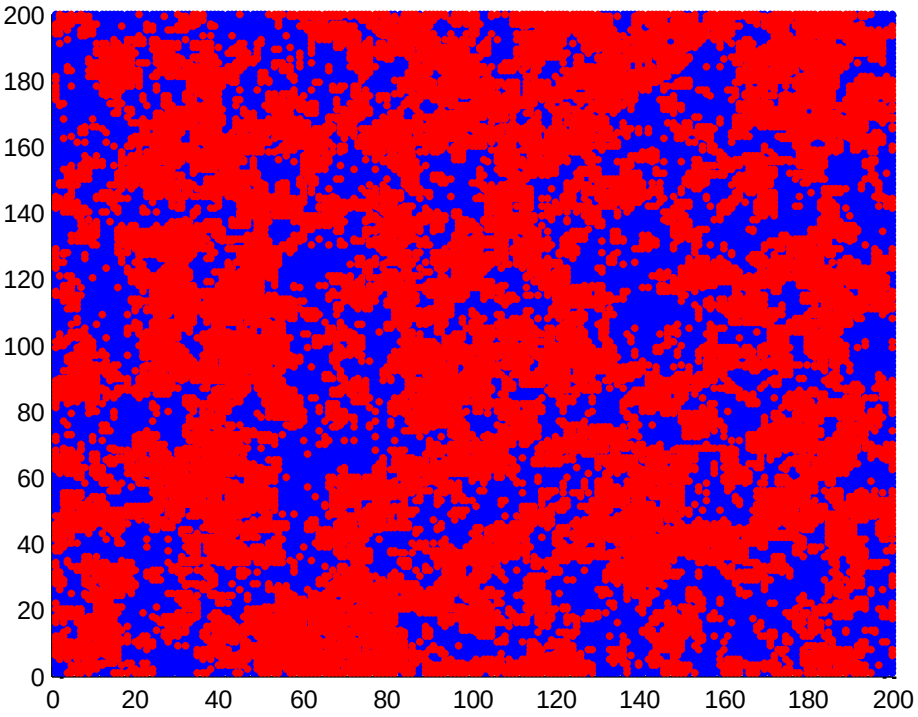


Figura 1.6. Modelo de Ising simulado em uma temperatura de T=2.6, com 1000 passos de Monte Carlo.

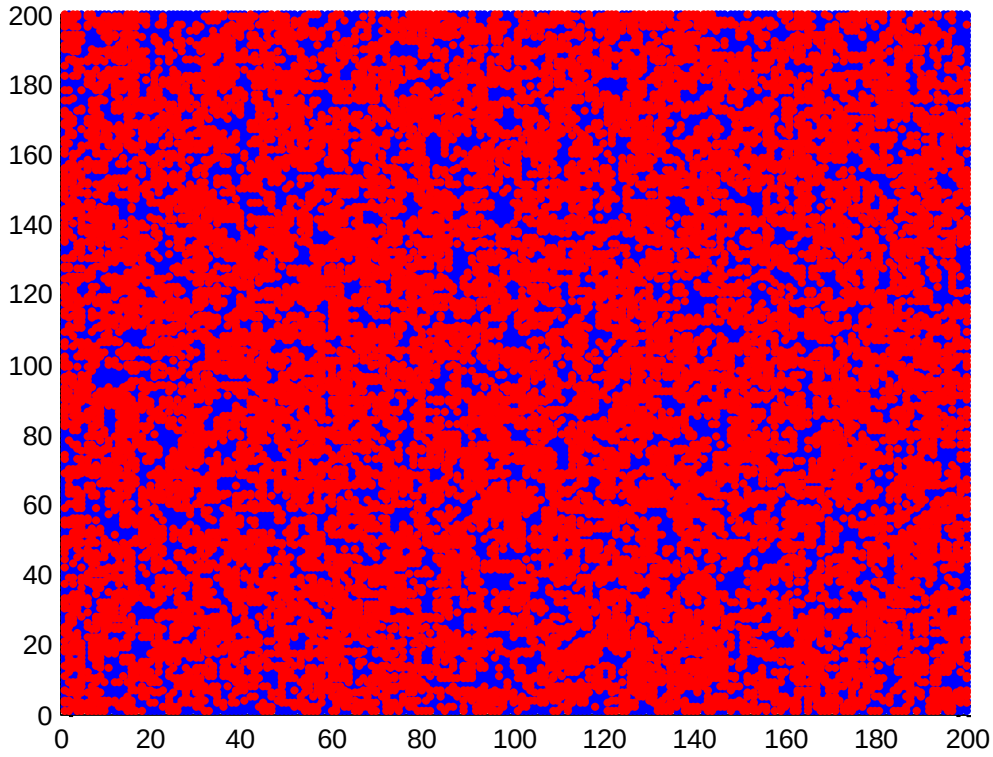


Figura 1.7. Modelo de Ising simulado em um a temperatura de T=4, com 1000 passos de Monte Carlo.

2 TÓPICOS DE FÍSICA ESTATÍSTICA

Neste capítulo será dada uma breve introdução baseado nos livros de Termodinâmica e Dinâmica Estocástica, Oliveira (2005) e Oliveira (2001), a respeito de fenômenos físicos como transição de fase, paramagnetismo, ferromagnetismo, superparamagnetismo, ensembles estatísticos e outros temas relacionados ao método de agrupamento de dado superparamagnético no ensemble microcanônico. Nas ultimas seções faremos um breve comentário sobre alguns métodos computacionais utilizados no desenvolvimento do programa, entre eles simulação no ensemble microcanônico e algoritmo de Hoshen-Kopelmann. Por fim descrevemos o método de agrupamento de dados superparamagnético no ensemble microcanônico.

2.1 FENÔMENOS FÍSICOS RELACIONADOS COM O MÉTODO EM ESTUDO

2.1.1 TRANSIÇÕES DE FASE

Transições de fase e fenômenos críticos ocorrem em uma enorme variedade de sistemas: fluidos simples e misturas de fluidos, materiais magnéticos, ligas metálicas e outros. Caracterizadas por uma não analiticidade dos potenciais termodinâmicos, refletida em divergências de suas derivadas que, como sabemos, estão diretamente relacionadas às respostas termodinâmicas (calor específico, susceptibilidade magnética e outras). Esse comportamento singular é causado por flutuações microscópicas que atingem e persistem em escalas macroscópicas, dando origem, portanto a um comportamento coletivo que fica bem caracterizado quando se investiga a correlação entre os constituintes do sistema. Nas vizinhanças do ponto crítico determinadas derivadas termodinâmicas, como a compressibilidade ou calor específico apresenta um crescimento sem limite.

A água, quando aquecida à pressão constante, entra em ebulição a uma temperatura bem definida transformando-se em vapor. Para cada valor de pressão a que está submetida a água, corresponde uma temperatura de transição. A temperatura de transição cresce com o aumento da pressão. Outros tipos de transições de fase ocorrem em física da matéria condensada, como em um material ferromagnético, por exemplo, onde a magnetização cresce à medida que

aumentamos o valor do campo aplicado. Por analogia a pressão está para o campo aplicado assim como a temperatura está para a magnetização do material. Podemos verificar outros tipos de transições em materiais ferromagnéticos, como a perda da imantação em uma temperatura bem definida, tornando-se, portanto em paramagnético. Os calores específicos e a suscetibilidade magnética apresentam um comportamento peculiar na região crítica, com divergências assintóticas que foram caracterizadas por meio de uma coleção de expoentes críticos.

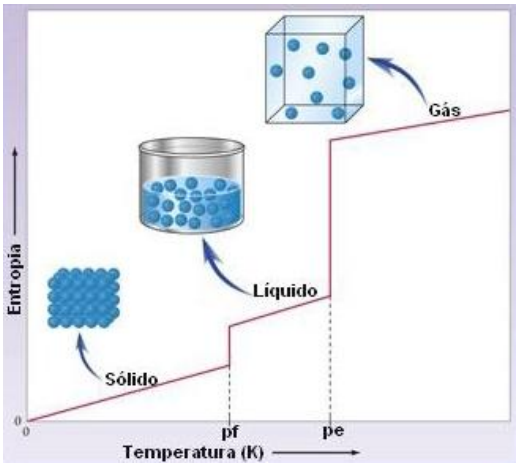


Figura 2.1 Gráfico da entropia em função da temperatura.

Na figura 2.1 vemos o fenômeno de transição de fase de uma substância observado através do gráfico de duas grandezas entropia em função da temperatura. Nos pontos de fusão e ebulição observamos que a temperatura permanece constante e a entropia cresce até a substância mudar de fase.

2.1.2 PARAMAGNETISMO

O estado paramagnético é caracterizado do ponto de vista macroscópico pela resposta linear a um campo magnético aplicado. Na ausência do campo, uma amostra de material não exibe magnetização. Aplicando-se um campo a amostra adquire uma magnetização que cresce linearmente à medida que a intensidade do campo for aumentando. Para pequenos valores do campo aplicado H, a magnetização m de uma amostra no estado paramagnético é proporcional ao campo, isto é,

$$m = \chi_o H \tag{2.1}$$

Em que, a susceptibilidade magnética χ_o, é positiva. Aumentando-se o campo, o comportamento de m com H deixa de ser linear e para valores do campo suficientemente altos, a magnetização satura atingindo um valor máximo.

A susceptibilidade χ_o depende da temperatura T. Para materiais que são paramagnéticos a todas as temperaturas, denominamos de paramagnetismo ideal, ela se comporta de acordo com a lei de Curie.

$$\chi_o = \frac{C}{T} \tag{2.2}$$

Em que C é uma constante positiva.

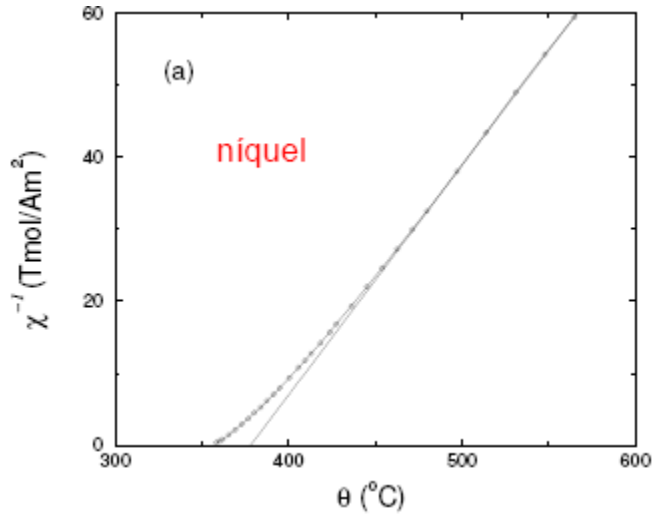


Figura 2.2 Susceptibilidade a campo nulo na fase paramagnética (extraída de Oliveira (2005)).

Em materiais que possuem uma fase paramagnética, mas não são paramagnetos ideais, a grandeza χ_o se comporta, para temperaturas suficientemente altas, de acordo com a lei de Curie-Weiss

$$\chi_o = \frac{C}{T - \Theta} \tag{2.3}$$

Sendo constante Θ positiva para materiais que sofrem transição para o estado ferromagnético. As duas constantes C e Θ podem ser determinadas ajustando uma assíntota aos dados experimentais de χ_o colocadas num gráfico de $1 / \chi_o$ versus T. A assíntota intercepta o eixo das temperaturas em T= Θ e a inclinação fornece $1 / C$. A constate Θ não deve ser confundida com a temperatura de transição.

2.1.3 FERROMAGNETISMO

As substâncias ferromagnéticas são caracterizadas por possuírem uma magnetização (espontânea) que pode persistir mesmo na ausência de campo magnético. Esse comportamento é bem diferente daquilo que ocorre numa substância paramagnética em que a magnetização desaparece quando o campo se anula. Se uma substância ferromagnética for aquecida as temperaturas suficientemente altas ela perderá a magnetização espontânea e se comportará como uma substância paramagnética, ou seja, ocorre uma transição de fase. Esta situação sugere que os *spins* dos átomos (ou moléculas) que constituem o material tenham uma forte tendência a se alinhar uns aos outros, dando origem a um momento magnético espontâneo. Na figura 2.3 ilustramos esquematicamente, o caso de uma pequena rede bi-dimensional.

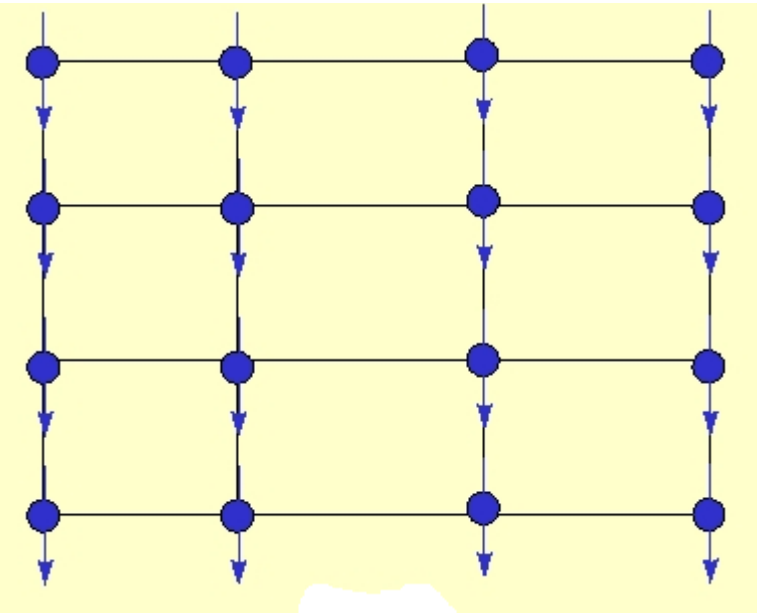
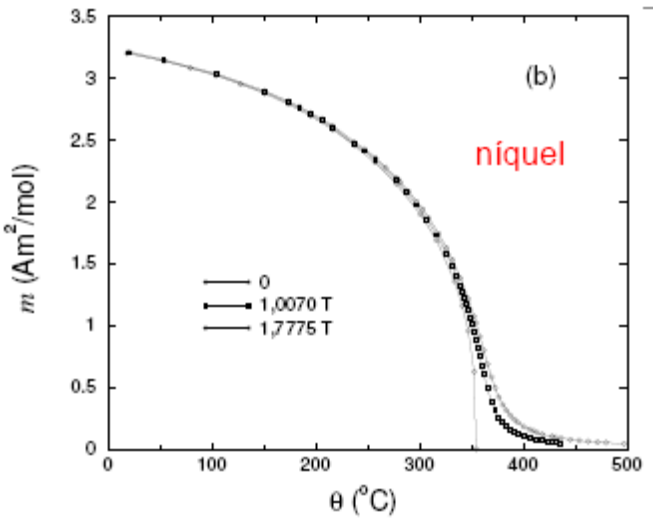


Figura 2.3 Representação da orientação dos Spins em uma rede regular.

As setas na figura 2.3 representam o *spin* do átomo (molécula). Esta orientação espontânea tende a desaparecer gradualmente à medida que o sistema é aquecido. Neste caso, os spins tendem a um estado de desordem. A temperatura crítica T_c é a temperatura na qual a magnetização espontânea desaparece, isto é, onde ocorre à transição entre "ordem" e "desordem".



Magnetização

Figura 2.4 Magnetização molar como função da temperatura para vários valores do campo aplicado (extraída do Oliveira (2005)).

Na figura 2.4 constatamos o efeito provocado pela temperatura sobre um material, à medida que cresce a temperatura a magnetização diminui até se anular para determinados valores.

2.1.4 SUPERPARAMAGNETISMO

O estado superparamagnético é caracterizado pela existência de regiões (clusters) do material ferromagneticamente ordenadas (magnetização não nula), porém descorrelacionadas entre si. Assim, o ordenamento do material como o todo é paramagnético. Pois o somatório da magnetização de todas as regiões, para campo magnético nulo, é igual a zero, o que caracteriza um fenômeno paramagnético. Em outras palavras podemos dizer que em altas temperaturas o comportamento macroscópico do sistema é dominado por flutuações aleatórias e o sistema encontra-se desordenado. Com a redução da temperatura pode ocorrer uma ou várias transições de fases dependo dos parâmetros que caracterizam o material.

2.2 TÓPICOS DE MECÂNICA ESTATÍSTICA

2.2.1 GRANDEZAS TERMODINÂMICAS

A capacidade térmica de um corpo é definida como a razão $Q / \Delta T$ entre o calor recebido Q e o correspondente incremento na temperatura ΔT. Mais precisamente, no limite desta razão quando ΔT→0. Sendo pequena a quantidade de calor, então Q=TΔS, em que ΔS é o incremento na entropia. Portanto, a capacidade térmica é dada por TdS/dT. Por definição $S = k_B \ln W$, onde W é o número de estados acessíveis ao sistema e k_B é a constante de Boltzmann.

A energia interna de uma amostra é definida como a média de ensemble das energias dos micro-estados,

$$U \equiv \langle E \rangle = \frac{\sum_m E_m \exp\left(-\frac{E_m}{k_B T}\right)}{\sum_{m'} \exp\left(-\frac{E_m}{k_B T}\right)} \tag{2.4}$$

Existe a seguinte relação entre a energia interna e a derivada logarítmica da função partição:

$$U = k_B T^2 \frac{\partial lnZ}{\partial T} \tag{2.5}$$

O calor específico, C_v, definido por

$$C_V = \frac{\partial U}{\partial T} \tag{2.6}$$

está relacionado com a variância da energia,

$$\sigma_E^2 = \langle E^2 \rangle - \langle E \rangle^2 \tag{2.7}$$

por

$$C_V = \frac{1}{k_B T^2} \sigma_E^2 \tag{2.8}$$

O parâmetro de ordem do sistema é a magnetização média dada por $\langle m \rangle$, o sistema quando desordenado apresenta magnetização igual a zero, para valores diferentes de zero há uma indicação de ordem magnética no sistema. Esta grandeza fornece uma idéia de alinhamento dos spins em uma dada direção. A magnetização, $m(S)$, associada à configuração dos spins é definida como:

$$m(S) = \frac{qN_{max}(S) - N}{(q-1)N} \tag{2.9}$$

Com $N_{max}(S) = max\{N_1(S), N_2(S), \dots, N_q(S)\}$, onde q é o estado de Potts apresentado pelo spin, N é o número de spins do sistema e $N_\mu\ (S)$ é o número dos spins com valor μ; $N_\mu(S) = \sum_i \delta_{s_i,\mu}$.

No estado superparamagnético o comportamento da magnetização indica o surgimento regiões com os spins alinhados, este fenômeno será usado no método para determinar a formação dos grupos.

2.2.2 ENSEMBLE E MÉDIA DE ENSEMBLE

Antes da definição propriamente dita definimos o termo ensemble segundo os conceitos de Mecânica Estatística, como um conjunto (assembléia) muito grande ($N \to \infty$) de sistemas idênticos (cópias mentais) à amostra, do ponto de vista macroscópico. A média de ensemble de uma variável A, denotada por $\langle A \rangle$ é a média aritmética dos valores de A sobre todos os sistemas do ensemble. Equivalentemente, se P_m é a probabilidade de encontrar o sistema (amostra) micro-estado m e A_m o valor de A neste micro-estado.

2.2.3 ENSEMBLE MICROCANÔNICO

No ensemble microcanônico o sistema de N moléculas, um volume V a energia é mantida constante em um intervalo entre E e $E + \delta E$, estas variáveis representam as únicas restrições possíveis existentes. Este sistema que pode ser gás, líquido e sólido é confinado em paredes adiabáticas, isto é, a energia não pode transmitida através das paredes existentes. A grandeza fundamental de ligação entre a termodinâmica e o ensemble microcanônico é a entropia. Não podemos colocar quaisquer outras limitações além daquelas que já foi imposto inicialmente, sendo permitidas as moléculas, fazer tudo que é possível com igual probabilidade de acesso aos diversos estados de energia E_r. Entre os estados microscópicos alguns são tão especiais que na realidade nunca serão acessados. Esta afirmativa pode ser representada pela expressão:

$$P_r = \begin{cases} W\ , se\ \ E < E_r < E + \delta E \\ 0,\ \ caso\ contrário. \end{cases} \tag{2.10}$$

O número total de estados microscópicos permitido em termos das limitações macroscópicas pode ser representado por W(E, δE, V, N). Assim, a probabilidade de

cada estado a ser acessado é igual a 1/W. Usando W, nós definimos a entropia S(E, δE, V, N) como segue:

$$S(\mathrm{E}, \delta\mathrm{E}, \mathrm{V}, \mathrm{N}) = \mathrm{k_b} \ln \mathrm{W}(\mathrm{E}, \delta\mathrm{E}, \mathrm{V}, \mathrm{N}) \tag{2.11}$$

Definimos a temperatura do sistema como:

$$T = \left[\frac{\partial S(\mathrm{E}, \delta\mathrm{E}, \mathrm{V}, \mathrm{N})}{\partial E}\right]^{-1} \tag{2.12}$$

Nesta equação a diferenciação parcial é feita mantendo se $\delta\mathrm{E}, \mathrm{V}, \mathrm{N}$ fixos. Para uma melhor compreensão da W, S, vamos investigar um exemplo simples de contagem de estados microscópicos. Temos aqui uma situação em que uma grande quantidade de dinheiro é distribuída entre uma população total de N pessoas. A quantidade de dinheiro e o número total de pessoas são variáveis macroscópicas, que são fixadas como restrições, isto é, eles são variáveis do estado. Se o dinheiro está sob a forma de uma barra de ouro, que pode ser dividida continuamente, e por isso seria difícil contar as formas de distribuí-la entre N pessoas. Esta situação é semelhante a que lidamos com um sistema da mecânica estatística. No entanto, no caso de dinheiro, há unidades como o real e seus centavos sendo assim quantizado como a energia. Se nós usamos essas unidades, o número de maneiras de distribuir o dinheiro é contável. O número de maneiras de distribuir E reais para N pessoas é dada como segue:

$$W = \frac{(E+N-1)!}{E!\,(N-1)!} \simeq \left(\frac{E+N}{E}\right)^{E} \left(\frac{N+E}{N}\right)^{N} \tag{2.13}$$

Para obter o resultado final da expressão acima usamos a fórmula de Stirling's. Assim a entropia e a temperatura são dadas como segue:

$$S(E,N) = k_b E \ln\left(1+\frac{N}{E}\right) + k_b N \ln\left(1+\frac{E}{N}\right) \tag{2.14}$$

e

$$\frac{1}{T} = \frac{\partial S}{\partial E} = k_b ln\left(1+\frac{N}{E}\right) \tag{2.15}$$

Expressando E em termos do parâmetro de temperatura T, nós obtemos:

$$E = \frac{N}{\exp\left(\frac{1}{\mathrm{k_b T}}\right) - 1} \tag{2.16}$$

A situação citada é análoga ao seguinte problema combinatório que pode ser vista em Salinas (2005), onde pretendemos distribuir M bolas idênticas dentro de N caixas dispostas ao longo de uma determinada direção. Para descobrir todas as

configurações possíveis, devemos calcular todas as permutações de $M + (N-1)!$ elementos (isto é, das bolas mais as divisórias que definem as caixas) e dividir o números obtido por $M!$ e $(N-1)!$ (pois as divisórias também são idênticas). Esta situação é idêntica ao exemplo acima a quantidade de bolas representa o número de reais que cada pessoa possui ou se pensarmos em um sistema de partículas a energia de cada uma.

2.3 MÉTODOS COMPUTACIONAIS APLICADOS NA IMPLEMENTAÇÃO

2.3.1 MÉTODO DO VIZINHO MAIS PRÓXIMO MÚTUO (KNN)

Pela sua simplicidade, é um poderoso método de classificação. Uma amostra é classificada de acordo com a sua proximidade com os k vizinhos, onde k é um inteiro. Para uma determinada amostra, geramos uma matriz de distâncias onde cada item é comparado com os outros itens do conjunto de dados. Estas distâncias são arranjadas em ordem decrescente. Elaboramos uma lista dos primeiros vizinhos de um dado item, em seguida propomos o critério de mutualidade para a lista de vizinhos. Dois elementos serão vizinhos se e somente se um estiver incluído na lista outro, caso isto não ocorra o elemento que foi consultado é excluído da lista. A escolha do k é essencial na execução do método. De fato, k pode ser considerado como um dos fatores mais importantes, gerando grande influência na qualidade das classificações, para um k muito pequeno vai levar a um número muito grande de variações nas classes. Por outro lado, a fixação em valores muito altos minimiza o número de classes, agregando elementos de classes distintas. Assim como qualquer outro parâmetro não existe um valor ótimo, em Bishop (1995) encontramos um algoritmo conhecido como validação cruzada que fornece uma estimativa para o número de vizinhos mais próximos.

2.3.2 ALGORITMO DE HOSHEN-KOPELMAN

Em muitos sistemas de simulação computacional, como por exemplo: percolação, autômatos celulares entre outros. Assim como em muitas outras aplicações científicas e comerciais (como contagem de bactérias em uma imagem digitalizada ou em microscópio óptico e reconhecimento de caracteres), estamos

interessados na possibilidade de localizar elementos conectados formando agrupamentos em um objeto em n-dimensão, que tenham alguma propriedade comum (como por exemplo, cor). Para estudar esta conectividade global seria necessário um número exagerado de operações. A fim de identificar e determinar o maior número de agrupamentos em um sistema observando se existe percolação, foi concebido um algoritmo muito rápido e é denominado Hoshen-Kopelman (1976). Os padrões são identificados com uma única varredura sobre os dados, reduzindo o tempo computacional no processo. Identificamos os agrupamentos através de cada linha da rede e rotulamos cada sitio que está ligado com seu vizinho mais próximo com um número. Assim, o agrupamento rotulado com $L_{i,j} = n$, onde n é o número que é atribuído ao agrupamento, cada sítio é inspecionado e verificado o vizinho superior e anterior quanto a sua similaridade com o sítio atual. Se existe similaridade e o número atribuído aos vizinhos é menor que o sítio verificado, então este sítio assume o rótulo de seu vizinho. Por fim avaliamos os rótulos e identificamos os elementos que estão conectados, dando aos elementos o menor valor entre os listados, identificando os agrupamentos existentes conforme a diversidade de índices.

A melhor maneira de descrever o algoritmo é através de um simples exemplo encontrado em Stauffer (2003). Suponha uma rede regular definida em duas dimensões com 15 células ocupadas são identificadas pela cor azul e as que estão vazias pela cor vermelha conforme a figura 2.5. As linhas são identificadas com um índice i e a colunas pelo índice j, o algoritmo faz a varredura das células são ao longo das linhas da esquerda para direita. O valor do rótulo na célula é definido conforme análise do rótulo do vizinho superior (i-1) e que está localizado à esquerda (j-1), a célula irá receber o menor rótulo estes dois vizinhos. Visitando a primeira linha rotulamos o primeiro elemento com 1, o segundo com 0 por não está ocupada. O terceiro com 2, o quarto com 0 por não está ocupada e o quinto com 3. Na segunda linha o primeiro elemento da linha é rotulado com 1, pois seu vizinho superior possui o rótulo 1. O segundo da segunda linha não existe ocupação recebe rótulo 0, no terceiro elemento rotulamos com 2 dado que o vizinho superior possui rótulo 2. No quarto elemento o rótulo dado é 2, pois o vizinho a esquerda tem rótulo 2. E no quinto observando o elemento superior vemos o rótulo 3 e o que está à esquerda possui rótulo 2, esta célula irá receber o menor valor neste caso 2, o elemento de rótulo 3 irá receber um rótulo impróprio, indicando que estes elementos

irão ser fundidos com os de rótulo 2. Ao observar a terceira linha o primeiro elemento recebe rótulo 1, pois o elemento superior tem rótulo 1, o segundo é rotulado com 1, pois o elemento da esquerda tem rótulo 1. O terceiro elemento da terceira linha possui o elemento superior com rótulo 2 e o elemento de esquerda com rótulo 1, será dado o menor rótulo e os elementos com rótulo 2 receberão um rótulo impróprio, sendo desta forma fundidos com os que têm rótulo 1. Na figura 2.6 podemos observar o processo de classificação descrito.

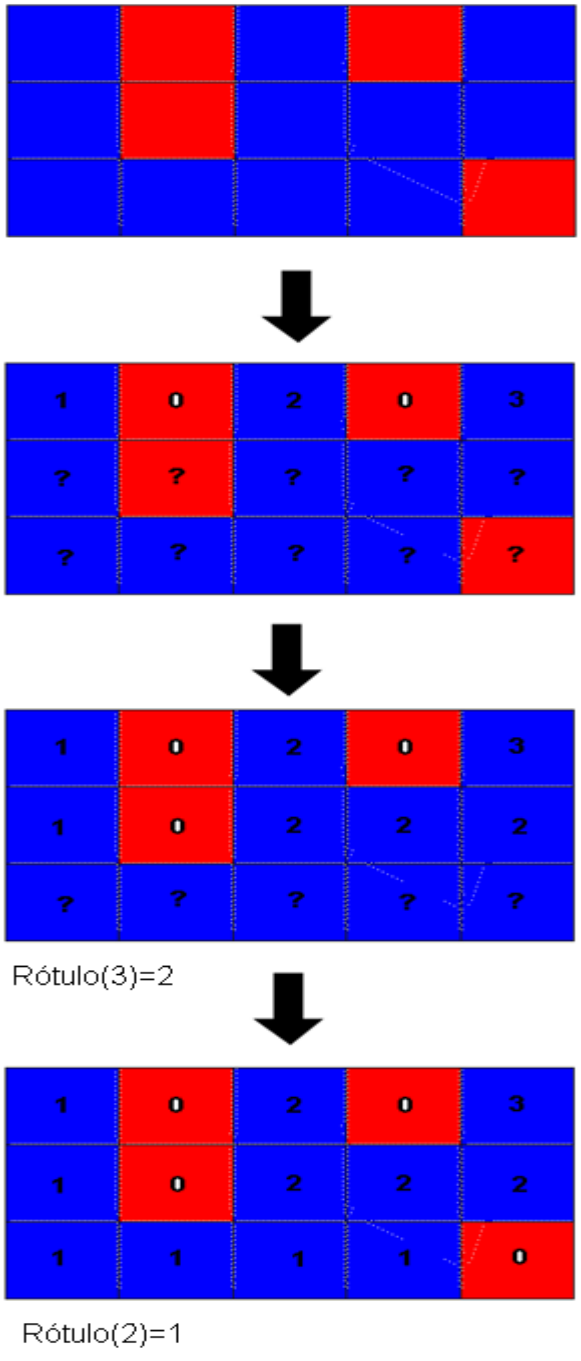


Figura 2.5. Processo de classificação segundo algoritmo de Hoshen-Kolpeman.

2.3.3 ALGORITMO DE SIMULAÇÃO NO ENSEMBLE MICROCANÔNICO

Uma alternativa para se evitar o custo computacional associado à geração de números aleatórios foi introduzido por Creutz (1983) um algoritmo que simula um ensemble microcanônico. No ensemble microcanônico o parâmetro de controle é a energia interna em lugar da temperatura como ocorre no ensemble canônico. Creutz imaginou a montagem do sistema através de variáveis aleatórias e mantendo as configurações de energia constante. A troca de energia no sistema está limitada a uma pequena partícula que transita ao longo dos contornos, absorvendo ou cedendo energia direcionando o sistema ao equilíbrio. Esta partícula consiste em um grau de liberdade extra, conhecida como demônio de Maxwell. Esta denominação teve origem na teoria do calor concebida por Maxwell em 1871. O que se denomina demônio de Maxwell é uma poderosa figura utópica, capaz de abrir e fechar um furo na fronteira de dois volumes contendo determinado gás à temperatura uniforme. O ser de poderes sobrenaturais é capaz de só deixar passar as moléculas mais rápidas em uma direção e as mais lentas na outra, como ilustrado na figura 2.7. Ele contraria o segundo princípio da termodinâmica, cria ordem a partir da desordem ou entropia inicial (velozes de um lado e lentos do outro) e, sem nenhum gasto de energia, pode elevar a temperatura de um lado e abaixa de outro. Levando o sistema ao equilíbrio podemos avaliar algumas propriedades como magnetização, calor especifico entre outras.

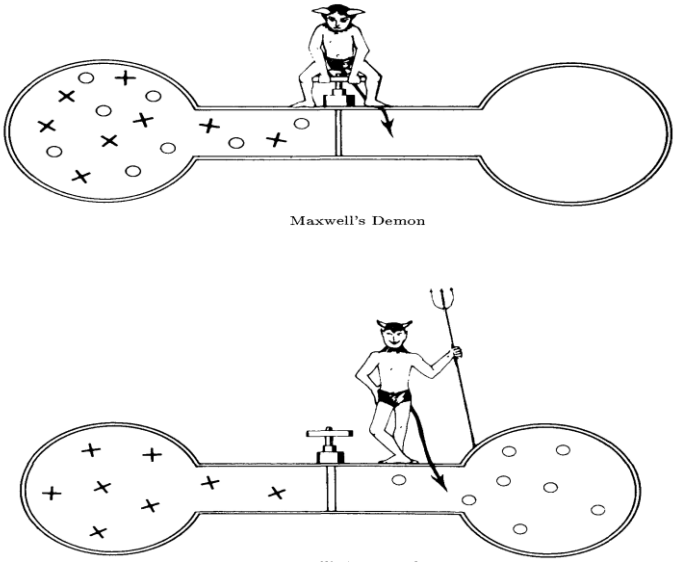


Figura 2.6 Ilustração do processo realizado pelo Demônio de Maxwell.

O método é mais facilmente descrito através de um exemplo. Tomemos, pois, o caso mais simples de um modelo de Ising numa rede regular. A energia configuracional do sistema é calculada através do hamiltoniano de Ising, no qual as partículas do sistema podem ser encontradas em dois estados, assumindo os valores $S_i = +1$ ou $S_i = -1$, e J é a energia de interação entre pares de spins vizinhos mais próximos. Em duas dimensões, cada spin interage com quatro spins vizinhos. A função do demônio é redistribuir a energia disponível ao sistema.

$$\mathcal{H}(\{S\}) = -J \sum_{\langle i,j \rangle} S_i S_j \,, \tag{2.17}$$

O procedimento computacional pode ser descrito nas seguintes etapas:

1) Construir uma rede aleatória;
2) Definir a energia do demônio Ed;
3) Sortear um sítio da rede;
4) Alterar a configuração gerando um novo estado local no sistema de modo que x → x';
5) Calcular a energia produzida na mudança, $\Delta\mathcal{H} = \mathcal{H}(x') - \mathcal{H}(x)$;
6) Se a energia é reduzida, aceitar a mudança, $Ed \longleftarrow Ed - \Delta\mathcal{H}$, substituir a atual configuração por x' e retornar para o passo 3;
7) Em caso contrário, aceitar a alteração apenas se o demônio possuir energia suficiente, isto é, $Ed - \Delta\mathcal{H} \geq 0$. Neste caso $Ed \longleftarrow Ed - \Delta\mathcal{H}$ e a nova configuração será x';
8) Retornar para o passo 3.

Conceitualmente vemos o demônio como um termômetro, em contato com as partes do sistema ganhando ou perdendo energia sucessivamente. Inicialmente a distribuição do demônio é arbitrária. O sistema funciona como um reservatório e um termalizador. Em última análise a energia do demônio segue a distribuição de Boltzmann, permitindo o cálculo da temperatura utilizado à expressão:

$$P(Ed) \propto exp\left(-\frac{Ed}{k_B T}\right) \tag{2.18}$$

2.3.4 DETALHAMENTO DO MÉTODO SUPERPARAMAGNÉTICO DE AGRUPAMENTO NO ENSEMBLE MICROCANÔNICO

O método consiste basicamente em identificar os clusters formados em pontos onde ocorrem transições de fase. As transições caracterizam se como uma ruptura nas ligações entre os clusters. Algo importante a ressaltar sobre as vantagens na aplicação do método superparamagnético no ensemble microcanônico é que o processo é não supervisionado, e o reconhecimento dos padrões não depende das formas geométricas do problema a formação de grupos emana do comportamento coletivo dos itens e é medido através de correlações entre os spins. O processo tem inicio com o cálculo das distâncias entre os itens da matriz de dados, foi usada no método a distância euclidiana, por ter a geometria idêntica a nossa situação real, porém o programa implementado permite o uso da Minkowski possibilitando variar em diferentes tipos de distâncias. Criamos então uma lista de vizinhos mais próximos de todos os itens da matriz. O próximo passo consiste em determinar a vizinhança mutua entre dos itens da lista. Através do algoritmo de Hoshen-Kopelmann o programa identifica os agrupamentos formados pelo método do vizinho mais próximo, e varia a quantidade de vizinho com o intuito de gerar um único cluster com um número mínimo de vizinhos, gerando uma espécie de bloco de itens inteiramente ligados. Para se ter um modelo com as propriedades de um imã granular não-homogêneo, é preciso obter as interações existentes entre os spins que serão fortes em regiões de altas densidades de dados e fracas interações onde a densidade é baixa. Utilizamos uma distância média local **a,** este valor é uma quantidade característica sobre a qual nossas interações crescem para altas densidades e decaem para baixas densidades. A expressão usada para calcularmos as interações J_{ij} entre os vizinhos é a seguinte:

$$J_{ij} = J_{ij} = \frac{1}{\hat{k}} \exp\left(-\frac{\left\|\overrightarrow{x_i} - \overrightarrow{x_j}\right\|}{2a^2}\right) \tag{2.19}$$

O valor $\hat{k}$ é a média de vizinhos por sítio. Note que o valor de **a** exerce uma importância maior sobre a função de interação que o $\hat{k}$, denotando a grande influência do comportamento coletivo no método. Sorteamos aleatoriamente os spins de Potts com o q assumindo valores que variam de 1 a 20 para cada item de dado,

criando assim um rotulo para os itens. Simulamos o sistema usando o método de Creutz, onde o demônio de Maxwell é um agente que distribui ou recebe energia do sistema, executando as trocas dos spins, isto é, do rotulo de cada dado, todo o processo ocorre em uma dinâmica de Monte Carlo.

O hamiltoniano é dado pela expressão:

$$\mathcal{H}[\{s\}] = -\sum_{\langle i,j \rangle} J_{ij}\delta_{s_i s_j} \qquad s_i = 1, \dots, q. \tag{2.20}$$

Onde $\delta_{s_i s_j} = 1$, se $s_i \neq s_j$ e $\delta_{s_i s_j} = 0$, se $s_i = s_j$. Os valores de $\mathcal{H}[\{s\}]$ determinam a energia suficiente para o demônio de Maxwell execute a trocar do estado de Potts q, ou receber a energia do sistema caso a nova configuração permita a doação para o demônio durante o processo.

Detalhadamente o método de Creutz consiste em sortear sítios e calcular o hamiltoniano e assim estimar a energia necessária para realiza a troca no spin. Ao estimar a energia o demônio é capaz de doar a quantidade necessária para executar a troca, em outro caso pode ocorre uma entrega de energia pelo item ao demônio e a troca é também efetuada. Ao longo de alguns passos de Monte Carlo o sistema encontra-se em equilíbrio, e podemos medir as grandezas relevantes para solução do problema.

Gerando um gráfico de magnetização versus energia podemos verificar as regiões onde ocorrem transições de fase, evidenciando assim uma ruptura dos blocos de itens que se encontrará ligado no começo do processo. Estas transições são representadas pelos patamares encontrados no gráfico, para realizar a análise dos agrupamentos formados simulamos o sistema em uma energia onde ocorre o patamar, através de critérios como a correlação spin-spin admitindo certo limiar, conseguimos por fim encontrar os agrupamentos formados, analisando aqueles que possuem uma grande correlação entre si. Verifiquem que este método de agrupamento de dados apresenta embutido no processo pelo menos três outros métodos como os vizinhos mais próximos é o Hoshen-Kopelman.

3 RESULTADOS E DISCUSSÕES

3.1 APLICAÇÕES

Neste capítulo faremos algumas aplicações para testar a eficiência do método proposto em diferentes bancos de dados. Por conveniência escolhemos exemplos amplamente estudados por outros métodos de agrupamento.

3.1.1 RUSPINI

O primeiro conjunto de dados é uma base conhecida como Ruspini (1970). Esta base de dados, representada na figura 3.1, foi gerada para testes de agrupamentos e é composta de 75 objetos bidimensionais, ou seja, cada objeto possui dois atributos. O conjunto de dados Ruspini é formado por quatro grupos com 20, 23, 17 e 15 elemento, respectivamente. Vamos detalhar a aplicação do método de agrupamento superparamagnético no ensemble microcanônico com este conjunto de dados.

Cada objeto representa uma entrada no conjunto de dados que é identificado por um número natural, que chamaremos de índice. A base Ruspini possui 75 itens. Logo, o primeiro receberá o índice 1, o segundo 2, o terceiro 3 e assim por diante até que o último receba 75. Precisamos construir um grafo completamente conectado, no qual a cada nó esteja associado um item e as arestas representem uma conexão entre os dados. Para que dois elementos estejam conectados, eles devem pertencer à vizinhança mútua. O grafo é construído iterativamente. Primeiro, calculamos a matriz de distância entre os itens. Nesta dissertação, empregamos a distância euclidiana, equação (1.4) com $\lambda = 2$. Em seguida, construímos uma lista com os k vizinhos mais próximos de cada item. Dois itens são vizinhos mútuos se o primeiro pertencer à lista do segundo e vice-versa. O algoritmo de Hoshen e Kopelman é, então, empregado para verificar a conectividade do grafo resultante. Durante o processo iterativo o programa parte com o número mútuo de vizinhos $k = 2$. O número de vizinhos é incrementado de um em um, até o valor necessário para formação de um grafo completamente conectado, como ilustrado na figura 3.2.

Identificadas as conexões entre os pares de itens, calculamos a intensidade de interação entre os spins de Potts correspondentes, dada pela equação (2.19). O problema está mapeado em um sistema magnético e pronto para se utilizar o algoritmo de Creutz. Ao i-ésimo nó do grafo associamos uma variável de Potts $s_i = 1, 2, \ldots, q$. Nesta aplicação utilizamos q=5. O hamiltoniano do sistema magnético granular é descrito pela equação (2.20).

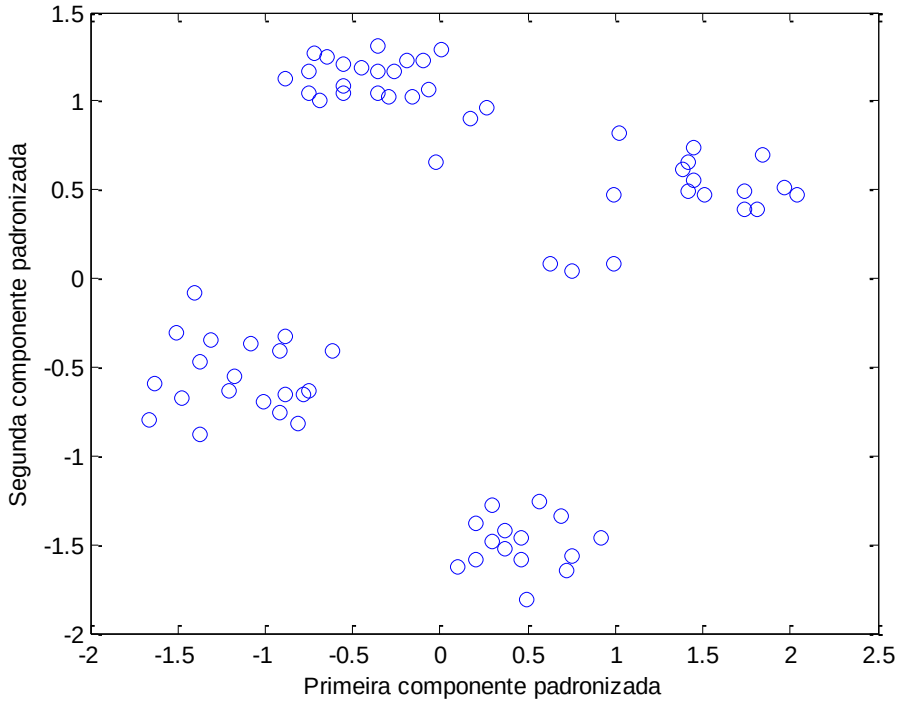


Figura 3.1 Conjunto de dados Ruspini.

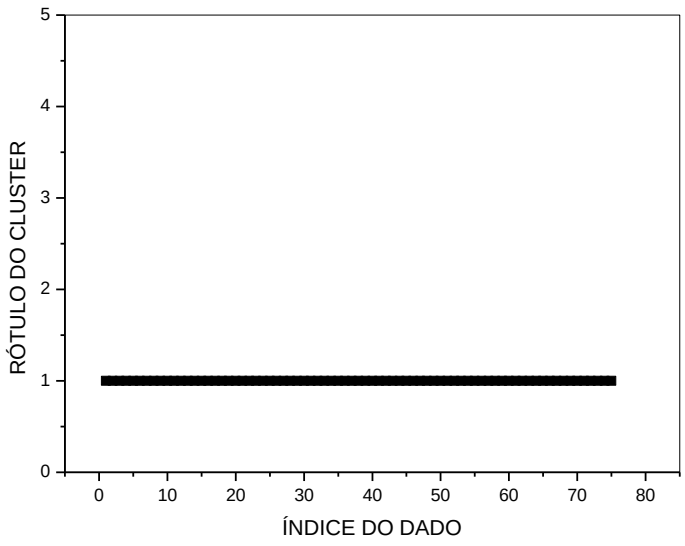


Figura 3.2 Base de dados após o processo de conexão total dos elementos.

As propriedades termodinâmicas do sistema são obtidas através do método proposto por Creutz, como descrito na secção 2.3.3 do capítulo 2. Especificamente, o demônio realiza um passeio aleatório através dos nós do grafo e tenta mudar o estado do spin de cada item visitado. O novo estado da variável de Potts é aleatoriamente selecionado entre as $q-1$ possibilidades com igual probabilidade. Caso a mudança diminua a energia do sistema, o novo estado é aceito e a energia liberada pelo sistema é absorvida pelo demônio. Por outro lado, caso a mudança aumente a energia do sistema, o novo estado só será aceito se o demônio possuir energia suficiente para fornecer ao sistema. Desta forma, a energia total do sistema composto é mantida constante e o vínculo sobre o demônio é satisfeito.

Após muitos passeios do demônio através do grafo, o sistema é levado ao equilíbrio e podemos calcular as grandezas termodinâmicas relevantes. Uma unidade de tempo corresponde a um passeio completo através do grafo.

Uma grandeza interessante para se monitorar ao longo da simulação é a magnetização total da amostra. A magnetização é considerada um parâmetro de ordem, pois ela é zero quando o sistema encontra-se completamente desordenado e diferente de zero quando o sistema apresenta alguma ordem magnética. Além disso, o valor da magnetização fornece uma idéia do tamanho da região ordenada.

Na figura 3.3 apresentamos a magnetização em função da energia para o magneto granular correspondente à base de dados Ruspini. Observamos que a magnetização é próxima do valor máximo para baixas energias e decresce sistematicamente à medida que a energia aumenta. A queda é gradual em certas faixas de energias, com quedas abruptas em energias específicas. Este comportamento indica que uma fração significativa dos spins torna-se descorrelacionados nestes valores de energia. Tais transições de fase demarcam a separação dos itens de dados em grupos, de acordo com Domany (1996). Neste caso, ocorrem quatro transições de fases indicadas pelas setas na figura 3.3. Antes que o sistema sofra a última transição, localizada em torno da energia -11, a magnetização é aproximadamente constante. Isto é uma indicação que os tamanhos dos grãos variam muito pouco para energias entre -19 e -11. Para valores de energia acima de -11, o sistema encontra-se totalmente fragmentado. Estes fenômenos ocorrem tanto para os dados padronizados quanto para os não padronizados. Para energias na faixa entre a penúltima e a última transição o

sistema é constituído por grãos de tamanhos relativamente estáveis, que devem corresponder aos grupos inerentes à base dados. Uma inspeção na figura 3.3, indica que em alguma energia em torno de -15 os grupos de itens de dados podem ser Identificados. Para identificar os grupos executamos o procedimento a seguir.

O sistema é simulado na energia identificada na etapa anterior e a correlação entre os vizinhos é medida. Um novo grafo é construído através da conexão entre os pontos cuja correlação é maior ou igual a certo limiar. Neste trabalho, utilizamos um valor de 0.5 para ativar uma ligação entre um par de vizinhos. Um grupo é formado pelos itens associados aos nós pertencentes ao mesmo subgrafo. A identificação dos elementos do grupo é realizada através do algoritmo de Hoshen e Kopelman, o qual atribui um mesmo rótulo aos nós que pertencem a um mesmo subgrafo. Nós isolados são posteriormente conectados ao nó vizinho de maior correlação.

As grandezas termodinâmicas devem ser estimadas no estado estacionário. Ou seja, após um número suficientemente grande de visitas pelo demônio. Para testarmos se o número de passos de Monte Carlo é bastante grande, monitoramos o comportamento das estimativas em função do número de visitas até que o erro estatístico se torne menor que uma tolerância pré-estabelecida. O efeito de um reduzido número de passos de Monte Carlo pode ser ilustrado na base de dados Ruspini. Aplicamos o método com os dados padronizados e sem padronização usando o parâmetro 1000 passos de Monte Carlo, ou seja, o demônio visita em média 1000 vezes cada nó do grafo. Na figura 3.3 podemos observar que o resultado com os dados padronizados apresenta menos flutuações que o resultado para os dados sem padronização.

Como vimos o algoritmo H-K atribui o mesmo rótulo a todos os nós de um mesmo subgrafo. Assim, a classificação dos dados pode ser visualmente ilustrada construindo-se um gráfico do rótulo atribuído ao nó, associado a um dado item do conjunto de dados, em função do índice que identifica o particular objeto no conjunto. Os gráficos da figura 3.6 resumem a classificação obtida para a base de dados Ruspini realizada com a energia igual a -15. Note que três itens são classificados incorretamente quando os dados estão sem padronização enquanto que dois são classificados incorretamente com os dados padronizados.

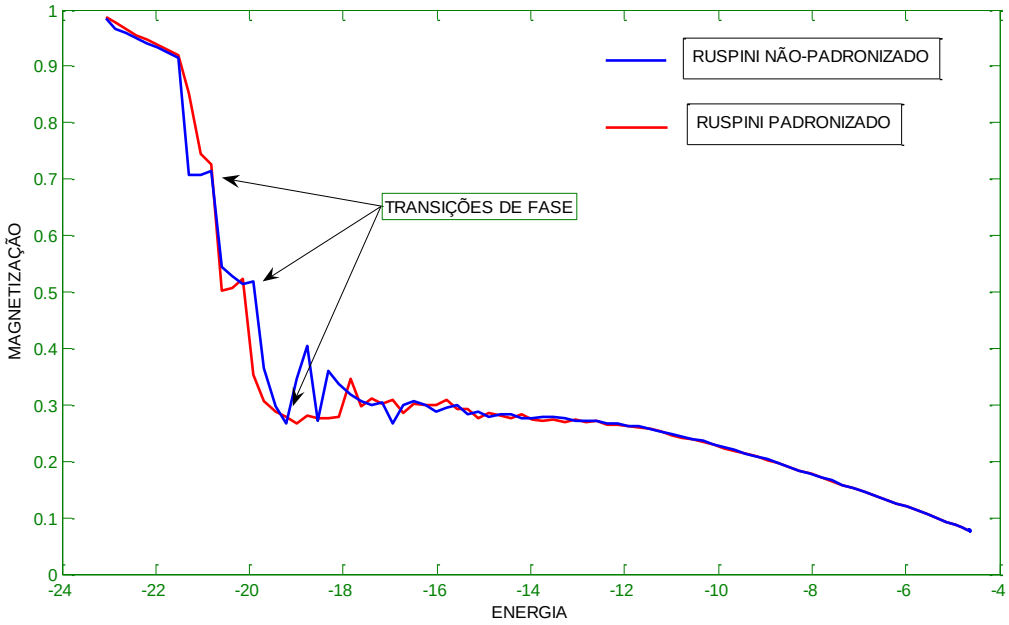


Figura 3.3 Magnetização em função de energia para os dados Ruspini. Algumas transições de fases estão assinaladas pelas setas e são caracterizadas pelas descontinuidades na magnetização.

Com o aumento do número de passos de Monte Carlo para 10000 com os dados padronizados, e o parâmetro de energia igual a -15 obtivemos a classificação representada na figura 3.7. Observe que 100% dos dados foram classificados corretamente.

Nas figuras 3.4 e 3.5 mostramos o comportamento termodinâmico do magneto granular associado à base de dados Ruspini em função da temperatura. Os gráficos apresentados evidenciam a ordem da transição, pois indicam a região do espaço de parâmetros onde ocorre coexistência de fases. Observe que para temperaturas em torno de 0.2 existem mais de um estado macroscópico para uma mesma temperatura.

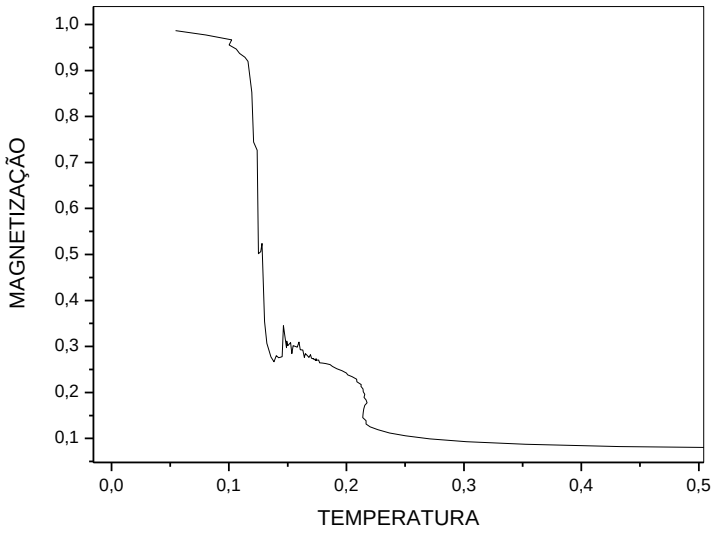


Figura 3.4 Magnetização em função da temperatura na base de dados Ruspini.

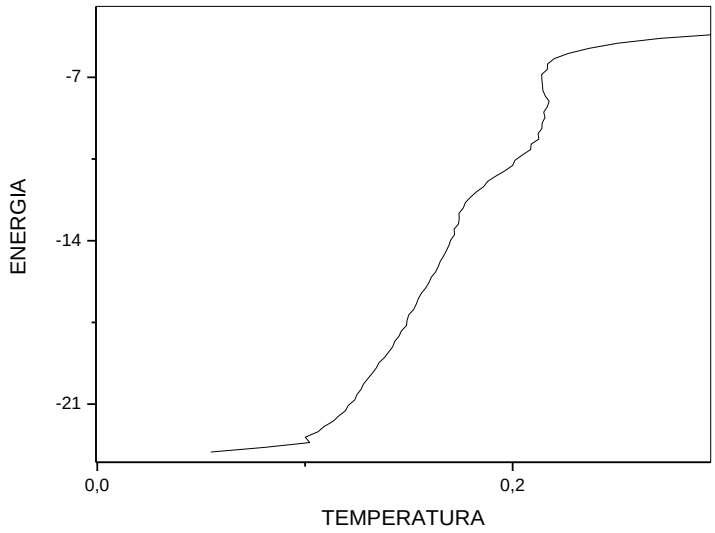


Figura 3.5 Energia em função da temperatura na base de dados Ruspini.

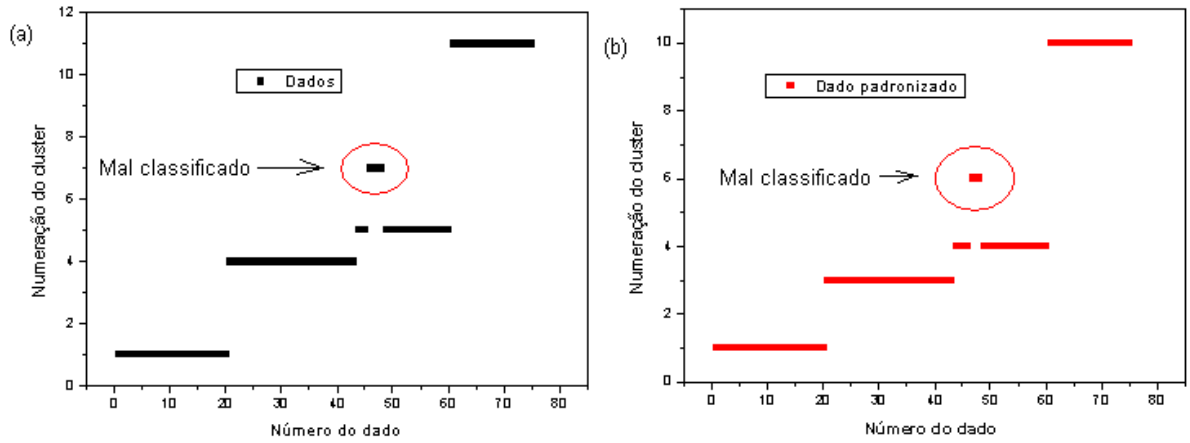


Figura 3.6 Resultados da classificação utilizando dados com e sem padronização em uma região de transição.

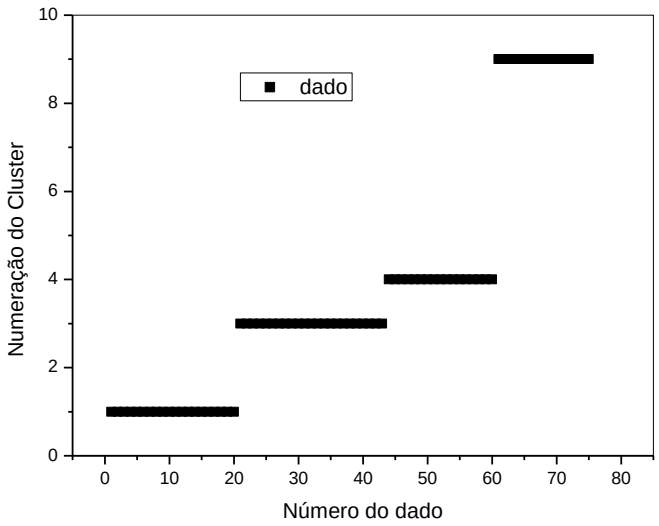


Figura 3.7 Resultados da classificação utilizando os dados padronizados em uma região de transição.

3.1.2 TRÊS RETÂNGULOS COM PONTOS DISTRIBUÍDOS UNIFORMEMENTE

O conjunto de dados apresentado nesta subsecção foi utilizado por Domany(1996) em uma das primeiras referência ao método de agrupamento superparamagnético. Os dados em questão consistem em três grupos com densidade uniforme, formando três retângulos. Cada região está longitudinalmente localizada nos intervalos $-20 < x < 20$ com $1.5 < y < 2.5$, $-0.5 < y < 0.5$ e $-2.5 < y < 2.5$, como pode ser observado na figura 3.8a. Cada retângulo possui 800 pontos e o ruído introduzido são representados por 800 pontos diluídos no intervalo $-3.5 < y < 3.5$ e $-35, < x < 35$, na figura 3.8b temos os dados padronizados representados graficamente. Realizamos o mapeamento em um magneto granular com o mesmo procedimento descrito na subsecção anterior. Através do método de Creutz, obtivemos a curva da magnetização em função da energia, mostrada na figura 3.9. Próximo à região assinalada em vermelho na figura 3.9, o sistema sofre a última transição de fase antes de se desintegrar. Medimos, portanto, a correlação entre os spins na energia -914, tomando média sobre 1000 configurações. Fomos capazes de identificar quatro grupos, com 865, 863, 528 e 349 elementos cada. Esta má classificação ilustra um aspecto importante da simulação computacional de sistemas de partículas fortemente interagentes. Como a correlação, e demais propriedades físicas, devem ser medidas no estado estacionário, é necessário que um número suficientemente grande de configurações seja desprezado antes de se tomar as médias. Além disso, as configurações sucessivas geradas pelo procedimento proposto por Creutz, por construção, não são estatisticamente independentes entre si. Isto implica que um número maior de configurações é necessário para reproduzir o comportamento de equilíbrio do sistema. E isto é mais relevante próximo a uma transição de fase. Em decorrência dessas observações, aumentamos o número de visitas do demônio para 2000 por nó do grafo. Com 2000 visitas por nó, identificamos os três grupos com 875, 839 e 875 elementos cada.

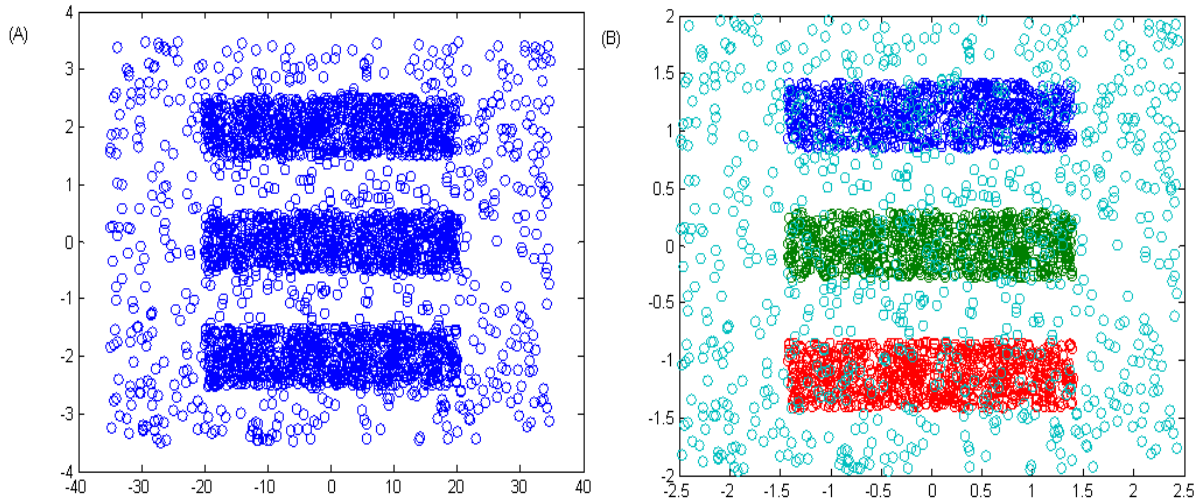


Figura 3.8 Representação dos pontos gerados no exemplo citado por Domany (1996) figura 3.8. A e os mesmos pontos padronizados figura 3.8.B cada retângulo possui 800 pontos distribuído uniformemente.

Os resultados obtidos pelo método de agrupamento superparamagnético no ensemble microcanônico representados na tabela 1 são muito próximos dos observados por Domany (1996) utilizando o ensemble canônico. Estes dados comprovam a eficiência do algoritmo pela proximidade dos resultados, em relação ao método superparamagnético realizado no ensemble canônico.

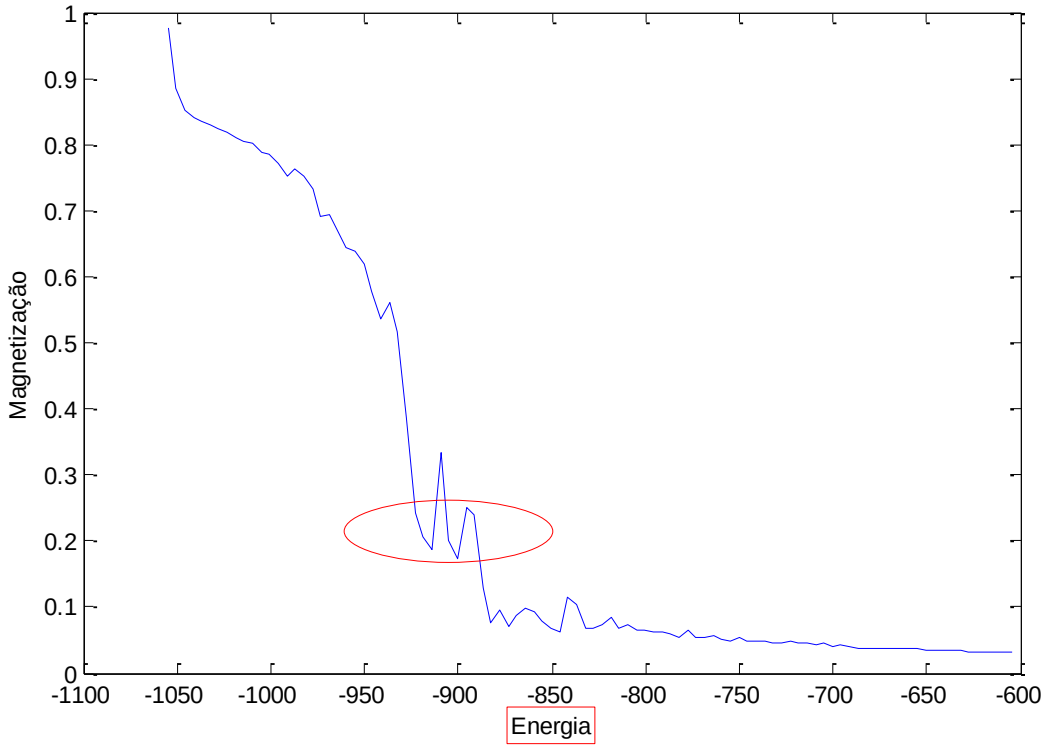


Figura 3.9 Gráfico da magnetização em função da energia, em destaque a região onde ocorre a transição de fase.

Tabela 1- Comparação entre os resultados do método de agrupamento superparamagnético no ensemble canônico Domany(1996) e no ensemble microcanônico.

Método	1º Retângulo	2º Retângulo	3º Retângulo
Canônico	883	874	863
Microcanônico	875	839	875
Valor exato	800	800	800

3.1.3 CONJUNTO DE DADOS ARTIFICIAIS EM TRÊS DIMENSÕES

Nesta aplicação, tomamos um banco de dados em três dimensões. Os dados correspondem a 1000 pontos sobre a superfície de um cilindro e 1000 pontos sobre a superfície de um toro. Os pontos estão aleatoriamente distribuídos sobre a superfície do objeto como mostrado na figura 3.10. Adicionamos ainda um ruído representado por 1000 pontos aleatoriamente distribuídos na região ocupada pelos objetos.

O raio da secção reta do toro é 0.1 e a circunferência que passa pelo centro desta possui um raio de 0.5. O raio do cilindro é 0.2 e seu corpo está delimitado pelos planos $z = 0.3$ e $z = 0.8$. Em seguida introduzimos o ruído diluído nos intervalos $-1 < x < 1$, $-1 < y < 1$ e $0 < z < 0.8$. Por fim todos os valores foram padronizados. Um banco de dados semelhante foi utilizado por Marangi(2001) numa implementação de uma técnica de agrupamento de dados que usa mapas caóticos.

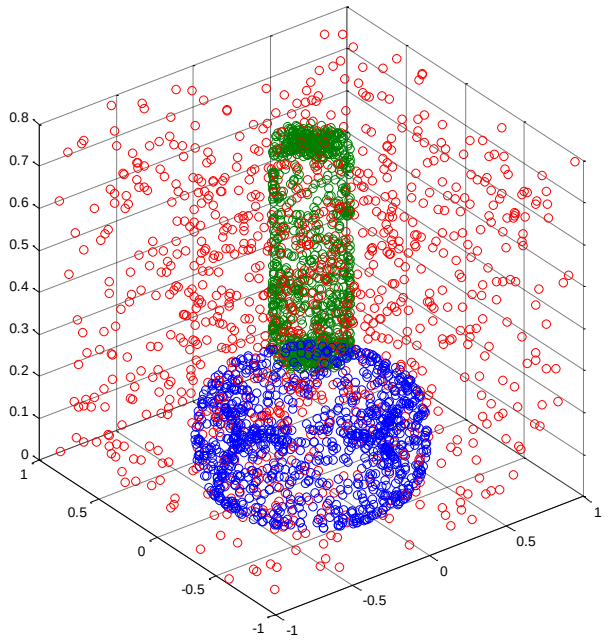


Figura 3.10 Duas superfícies tridimensionais um cilindro e um toro, imersos em pontos distribuídos na região onde as figuras estão localizadas.

A figura 3.11 mostra o comportamento da magnetização em função da energia para este sistema. Em destaque, a região onde ocorre a transição de fase.

Podemos observar uma larga faixa de energias onde a magnetização é aproximadamente constante. Para energias nesta faixa, certos spins estão temporalmente fortemente correlacionados, ou seja, em grupo. Medimos a função de correlação spin-spin na energia -840, aproximadamente no ponto médio do plateau, com 3000 varreduras por spin. Identificamos a formação de dois grandes agrupamentos com 1036 e 1069 objetos respectivamente. Os pontos devido ao ruído forma vários grupos de pequenos. Na figura 3.12, mostramos os pontos pertencentes a cada um dos dois agrupamentos formados. Note que o ruído foi quase completamente eliminado. Este resultado é superior ao apresentado por Marangi(2001) no mesmo sistema e demonstra a capacidade da técnica de agrupamento superparamagnético no ensemble microcanônico de filtrar ruídos de imagens.

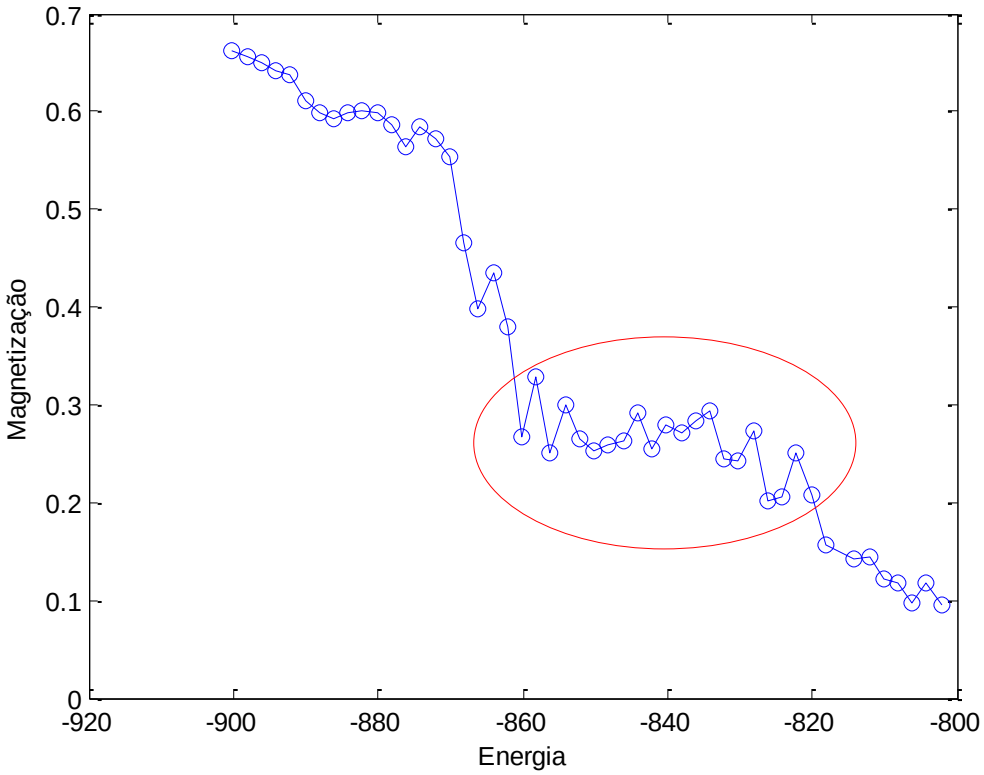


Figura 3.11 Gráfico da magnetização em função da temperatura em destaque a região em que ocorre a transição de fase.

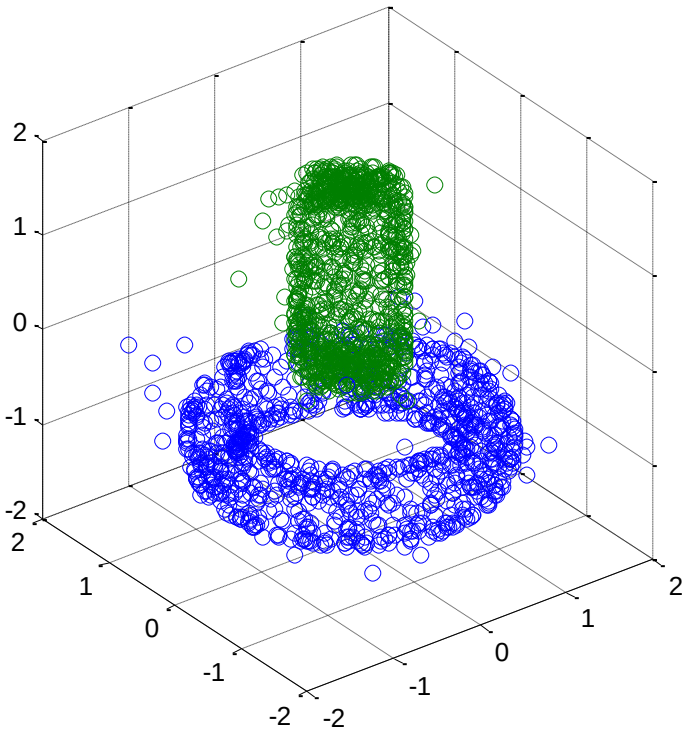


Figura 3.12 Resultados da classificação do cilindro e do toro obtido através do método de agrupamento.

Na tabela 2 podemos comparar os resultados da classificação dos objetos representados em três dimensões utilizando o método de agrupamento superparamagnético no ensemble microcanônico e os obtidos utilizando o método de agrupamento por mapas caóticos Marangi (2001). Percebemos que a eficiência do método superparamagnético é bastante superior e os resultados estão muito próximos aos valores exatos.

Tabela 2- Comparação entre resultados obtidos pelo método de agrupamento superparamagnético no ensemble microcanônico com o método de agrupamentos por mapas caóticos.

Método de agrupamento	Tiróide	Cilindro
Mapas caóticos	1221	1237
Microcanônico	1036	1069
Valor exato de pontos	1000	1000

3.1.4 TRÊS SUBESPÉCIES DA FLOR ÍRIS

Por último vamos aplicar o procedimento de agrupamento descrito nesta dissertação a um caso real. Trata-se de um exemplo clássico apresentado por R. A. Fisher (1936) e utilizado freqüentemente como padrão de comparação entre métodos de agrupamentos e aferição de eficácia. O conjunto de pontos para agrupamentos consiste em 150 amostras de quatro características métricas de flores do tipo Íris. Logo, os pontos são elementos de R^4, isto é, espaço euclidiano de dimensão 4. Cada ponto consiste em comprimento e largura da sépala, comprimento e largura da pétala, e foram extraídas amostras de três grupos distintos de 50 espécimes cada (50 de Setosa, 50 de Versicolor e 50 de Virgínica) que podem ser vista na figura 3.13. Todos os dados deste exemplo foram padronizados de maneira que nenhum atributo tenha maior relevância sobre o outro.

Figura 3.13 Foto das três espécies da planta íris à esquerda Íris Versicolor, no centro a Íris Setosa e a direita Íris Virginica.

Nas figuras 3.14 e 3.15 mostramos a projeção dos dados em três dimensões. Na figura 3.14, os atributos estão com as medidas originais e na figura 3.15 as medidas foram padronizadas. Note que duas das três espécies estão aglomeradas numa região bem distante da região ocupada pela terceira.

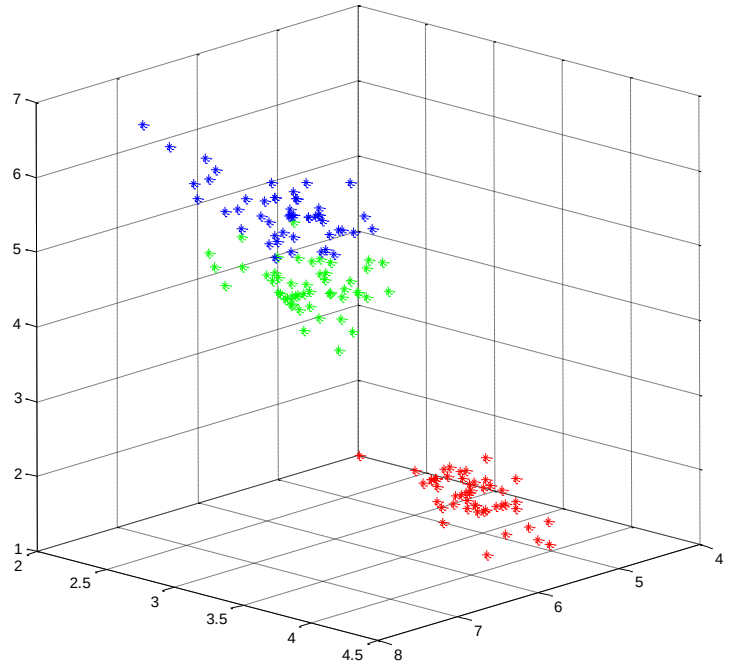


Figura 3.14 Representação dos pontos através de três atributos, os pontos em vermelho representam a Íris Setosa, em verde e vermelho as íris Versicolor e virginica respectivamente.

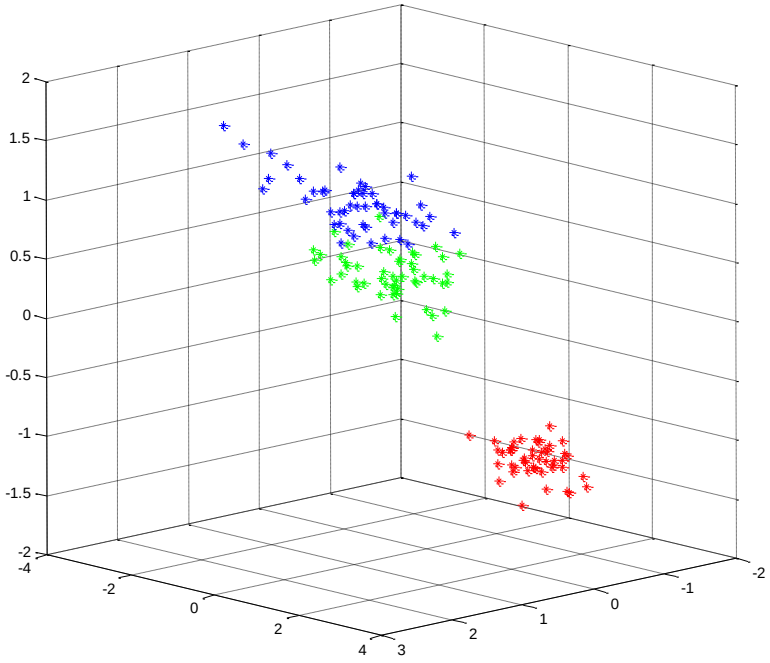


Figura 3.15 Representação dos pontos do banco de dados padronizados da planta íris através de três atributos.

Nesta aplicação usamos o mesmo procedimento dos casos descritos anteriormente. Inicialmente, simulamos o sistema de spins com os parâmetros sugeridos por Domany (1997). A saber, o número de vizinhos mútuos $k = 5$ e spins de Potts com $q = 20$ estados. A curva da magnetização em função da temperatura para o magneto granular correspondente à planta Íris é mostrado na figura 3.16. Da curva de magnetização, parece evidente que o sistema sofre duas quebras, antes de a magnetização apresentar algumas flutuações para energias acima de -41.

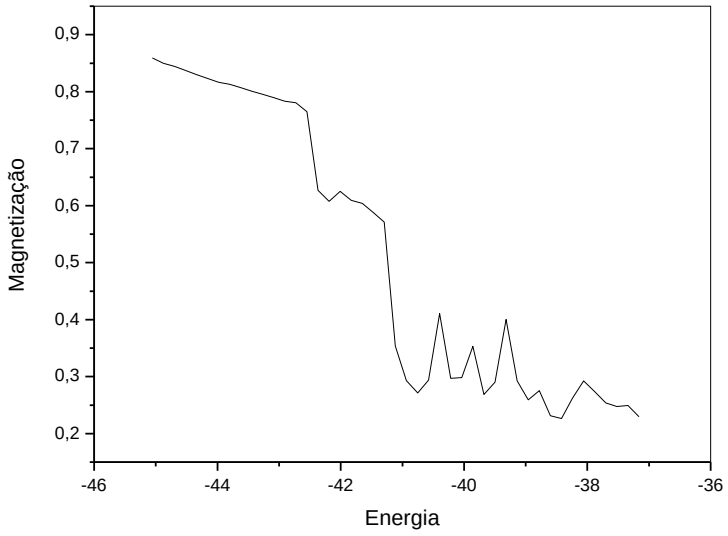


Figura 3.16 Gráfico da magnetização em função da energia para o banco de dados da planta íris.

De fato, verificamos uma transição de fase ocorrendo em um valor de energia -42.3. Uma tentativa de realizar o agrupamento neste valor de energia fornece dois grandes agrupamentos um com 50 e o outro com 79 elementos. Pudemos verificar que o grupo de 50 elementos é formado pelos exemplares Íris setosa, enquanto que o outro possui exemplares das duas outras espécies. O comportamento da magnetização sugere uma transição de fase em -40.3, onde esperamos ocorrer a separação das outras espécies em dois grupos. A análise via função de correlação spin-spin nesta energia identificou os três maiores agrupamentos com 47, 58 e 21 elementos. Após a classificação dos grupos, 102 elementos foram classificados corretamente como podemos perceber no gráfico da figura 3.17. Os elementos que estão na fronteira ou um pouco mais distantes do grupo do qual faz parte não foi

bem classificado. O alto estado de agregação dos itens na região de fronteira faz com que haja uma maior interação entre os elementos de fronteira em relação aos que estão mais afastados. Esta observação nos leva a afirmar que a energia necessária para romper uma ligação na fronteira é maior que a energia de ligação dos elementos mais afastados. Ao adicionar uma pequena quantidade de energia ao sistema acabamos rompendo ligações de elementos do mesmo grupo e mantendo as de grupos distintos.

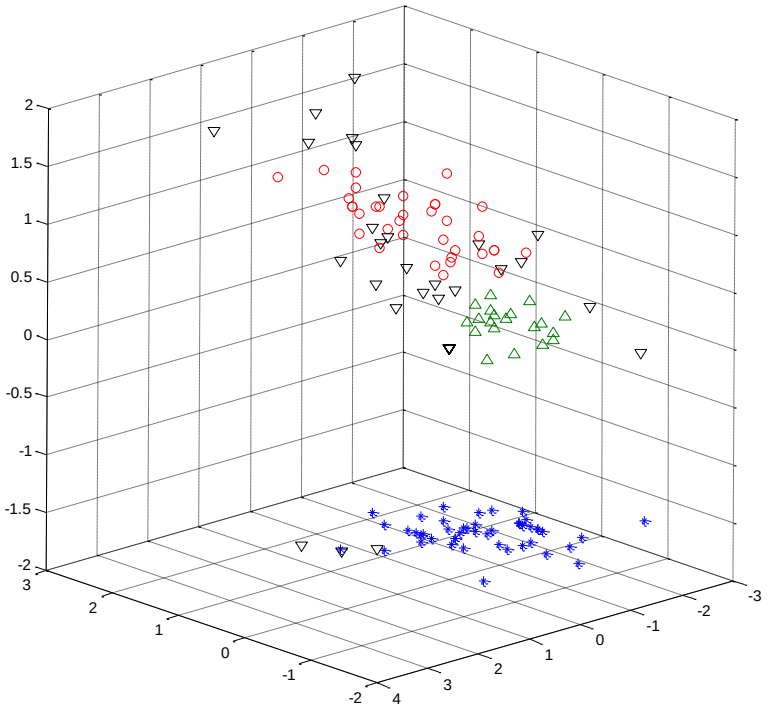


Figura 3.17 Elementos classificados corretamente nos dados da planta íris. Em um total de 150 elementos, 102 foram classificados corretamente, traduzindo uma eficiência de 68%.

Na tabela 3 temos uma comparação entre os resultados obtidos pelo método de agrupamento superamagnético nos ensemble canônico Domany (1996) e no ensemble microcanônico, observamos uma insuficiência do método para ensemble microcanônico, as duas espécies que possuem os agrupamentos muito próximos contribuíram significativamente para este insucesso.

Tabela 3- Comparação dos resultados do método de agrupamento Superparamagnético nos ensembles canônico e microcanônico, para os dados da planta Íris.

Ensemble	Íris Setosa	Íris Versicolor	Íris Virginica
Canônico	38	45	40
Microcanônico	47	58	21
Valor exato	50	50	50

A identificação dos grupos do banco de dados da planta Íris é reconhecidamente um desafio para os algoritmos de agrupamento. Não obstante, classificar corretamente 102 elementos em um total de 150 é um pouco abaixo do resultado apresentado por Domany (1996), com a simulação realizada no ensemble canônico. Vamos investigar um pouco mais como método de agrupamento superparamagnético no ensemble microcanônico se comporta quando os parâmetros do modelo são variados.

Primeiro, testamos o efeito de aumentarmos o número de vizinhos mútuos. A ampliação da vizinhança mútua não provocou mudanças significativas nos grupos identificados. Porém, as curvas das grandezas físicas apresentaram menos flutuações. Por outro lado, o número de estados de Potts, q, muda drasticamente o comportamento termodinâmico do sistema. É interessante observar na figura 3.16 que a faixa de energias que começa logo após a segunda transição de fase é muito estreita. Uma tentativa de classificação em energias acima de -41 resulta em um número muito grande de grupos. Além disso, a classificação resultante não tem reprodutibilidade e é fortemente influenciada pelo tempo de simulação. Isto é um forte indício de que o sistema é termodinamicamente instável nesta região.

O quadro acima descrito pode ser testado através do comportamento da energia em função da temperatura. Próximo a uma transição de fase de primeira ordem existe uma região de coexistência onde duas ou mais fases distintas podem estar presentes. Ou seja, o sistema pode ser encontrado em mais de um estado macroscópico para um mesmo valor da variável de controle. Isto é refletido no comportamento da energia em função da temperatura que apresenta uma curva em forma de S, como mostrado na figura 3.18. Note que a forma em S se torna mais acentuada à medida que o número de estados do spin cresce. A derivada da energia

em relação à temperatura corresponde ao calor o específico, que deve ser positivamente definido devido à segunda lei da termodinâmica. Assim os estados macroscópicos associados à região de derivada negativa na figura 3.18 são termodinamicamente instáveis. E podemos observar que a região de instabilidade cresce à medida que aumentamos o número de estados de Potts.

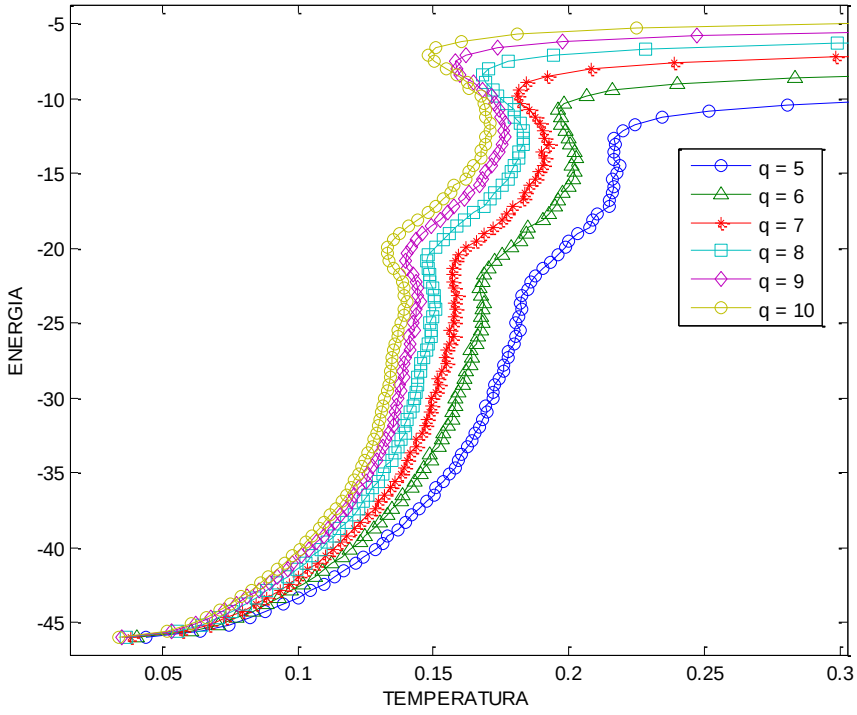


Figura 3.18 Energia em função da temperatura para diferentes números de estados de Potts.

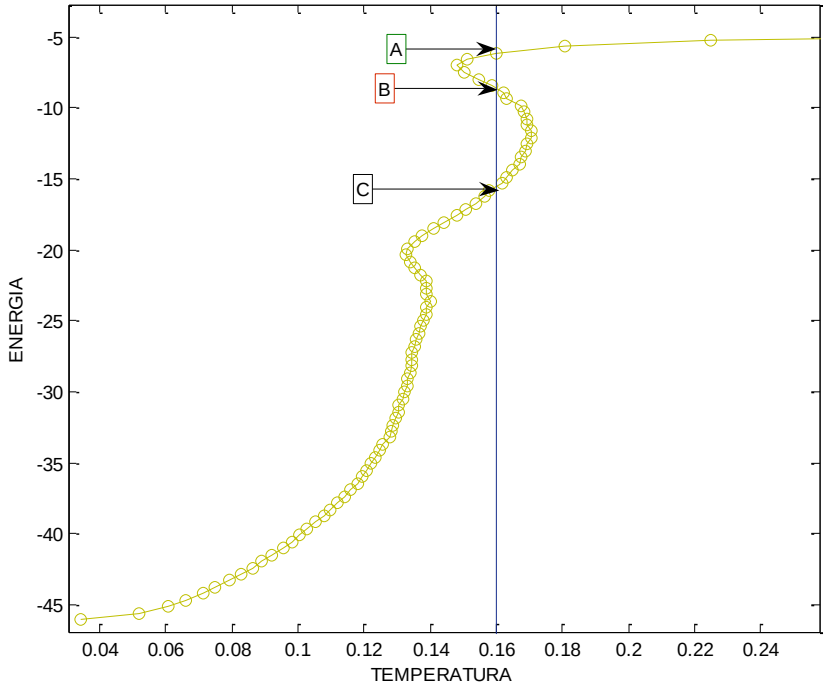


Figura 3.19 O gráfico mostra os diferentes estados macroscópicos para uma mesma temperatura.

Na figura 3.19, destacamos o comportamento da energia com a temperatura para um sistema de Potts com 10 estados correspondentes aos dados da planta Íris. Observamos que para uma mesma temperatura existem três possíveis estados macroscópicos nos o sistema poderia ser encontrado, indicados pelas setas A, B e C. Os estados A e C são metaestáveis, e o sistema pode ser encontrado em qualquer dos dois estados. Enquanto que os estado B é instável, pois a derivada neste ponto é negativa. No ensemble canônico o sistema transita entre os estados A e C no decurso da simulação. Ou seja, o sistema salta de ida e volta entre microestados característicos do macroestado A e B. Isto compromete a confiabilidade dos valores médios das propriedades do sistema obtidos na simulação no ensemble canônico.

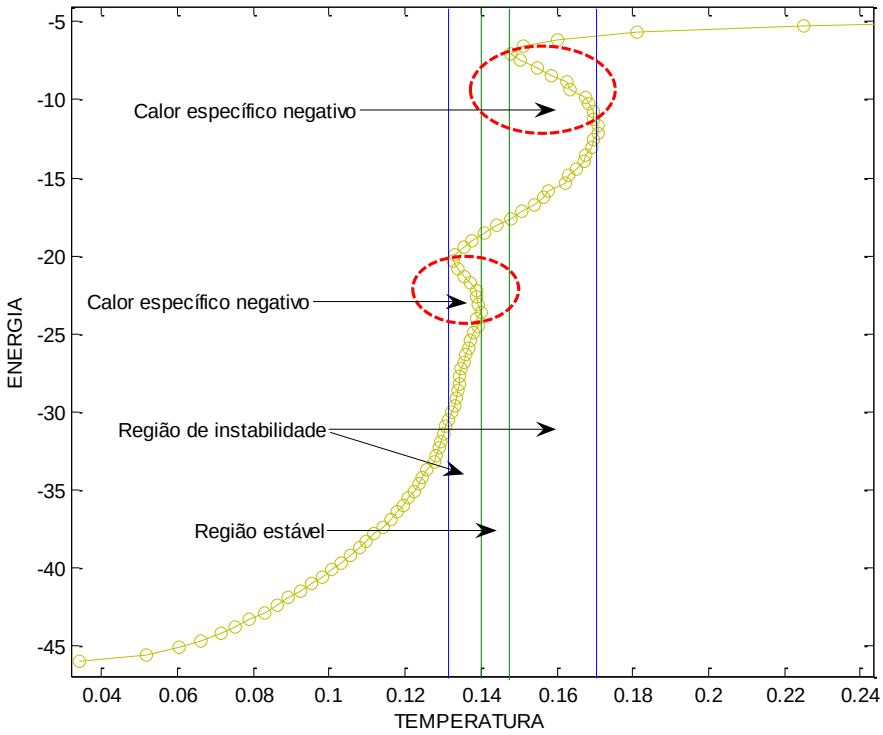


Figura 3.20 Regiões de instabilidade no sistema.

Na figura 3.20 verificamos que existem duas regiões de instabilidade das quais não podemos utilizar os valores de energia, principalmente nos locais onde o calor específico é negativo, violando desta forma as leis da termodinâmica. Na região estável a quantidade de energia aplicada ao sistema é muito superior ao valor necessário para quebrar os dois grupos, ocorrendo o rompimento de vários outros elementos que se encontram em um estado de interação com a energia bem menor. No gráfico de figura 3.21 vemos a região onde supostamente deveria haver uma transição de fase e é possível constatar que nesta região a instabilidade do sistema é muito alta gerando problemas na classificação correta para o banco de dados da planta íris. Como foi dito anteriormente esta região cresce com o aumento do número de estados de Potts para q=20 não existe região estável para a classificação.

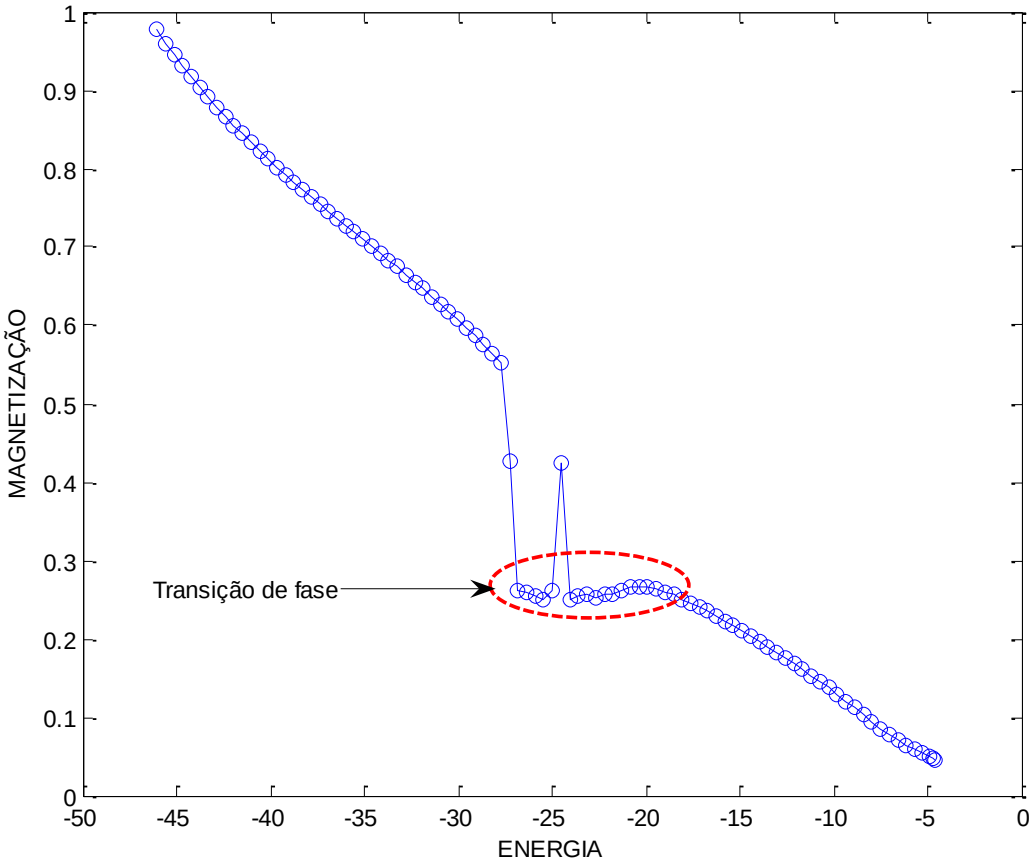


Figura 3.21 Gráfico magnetização em função da energia para q=10.

CAPÍTULO 4

CONCLUSÕES

O método de agrupamento de dados superparamagnético, implementado em um ensemble microcanônico apresentou excelente desempenho na identificação de grupos em duas e três dimensões. Em padrões como o Ruspini e nas superfícies em três dimensões, observamos um índice de acerto acima de 90%. No banco de dados que consistia em três retângulos os resultados foram semelhantes aos obtidos com o método superparamagnético no ensemble canônico. Ressaltamos que no ensemble microcanônico o tempo de execução do programa é bem menor, pois não precisamos gerar tantos números aleatórios e o algoritmo é muito simples. Em particular, o método é adequado para lidar com dados contaminados por ruídos. Imagens em duas dimensões apresentam maior dificuldade em filtrar ruídos, pois os pontos de dados e os do ruído estão no plano. Com o aumenta do número de dimensões a incidência de ruídos no conjunto de elementos classificados diminui bastante.

O único conjunto de dados reais tratado nesta dissertação foi o bem conhecido banco de dados da planta Íris, exaustivamente utilizado como teste para novas propostas de métodos de agrupamento. O desempenho de nossa implementação foi razoável, porém inferior ao reportado por Domany e colaboradores (1996) através do mesmo método implementado no ensemble canônico. Uma investigação detalhada das propriedades termodinâmicas do sistema físico no qual os dados são mapeados revelou que a multiplicidade de fases coexistentes pode dificultar sobremaneira o bom desempenho do método. Em particular, verificamos que para certos valores de número de vizinhos mútuos e grandes números de estados de Potts não há uma fase termodinamicamente estável que permita a correta classificação, pois os tamanhos dos “grãos” magnéticos flutuam.

A discussão a respeito dos pobres resultados no banco de dados da planta Íris foi relevante para avaliar as verdadeiras potencialidades do método e quanto à chance de falha na classificação. Isto confere um teste de auto-consistência à

proposta de agrupamento não supervisionada, baseada no comportamento termodinâmico de um magneto não-homogêneo. Verificamos a partir deste banco de dados que para algumas situações é necessário observar o comportamento da energia em função da temperatura e não apenas localizar as transições de fase. Pois devemos evitar realizar classificações em regiões de instabilidade. Desta forma poderemos prever antecipadamente se o método apresentará resultados confiáveis, esta afirmação faz com que este método se diferencie dos outros.

REFERÊNCIAS BIBLIOGRÁFICAS

BISHOP, C. M. Neural networks for pattern recognition (Clarendon Press, Oxford, 1995).

CREUTZ, M. Microcanonical Monte Carlo simulation, Phys. Rev. Lett. 50, 1411-1414 (1983).

ANGELINI, L. Clustering data by inhomogeneous chaotic map lattices, Phys. Rev. Lett. 85, pp. 554-557, (2000).

MARANGI, C. Clustering by inhomogeneous chaotic maps in landmine detection, Proc. SPIE vol. 4170, (2001)

DOMANY, E. Advances in Neural Information Processing Systems 8, 416 (1996).

DOMANY, E. Super-paramagnetic clustering of data - the definitive solution of an ill-posed problem, Progress in Statistical Physics, Proc. Choh Memorial Conference, Seoul 1997; World Scientific, p. 213 (1998).

DOMANY, E. Super-paramagnetic clustering of data, Phys. Rev. E 57, 3767 (1998).

DOMANY, E. Super-paramagnetic clustering of data, Physical Review Letters 76, 3251 (1996).

DOMANY, E. Super-paramagnetic clustering of data, Proc. STATPHYS20, Paris 1998; Physica A 263, 158 (1999).

DOMANY, E. Super-paramagnetic clustering of yeast gene expression profiles, Physica A279, 457 (2000).

DOMANY. E. Super-paramagnetic clustering of data: application to computer vision, Conference on Computational Physics, Granada, 1998; Comp. Phys. Comm. 121-122, 5 (1999).

DORIGO, M.; STÜTZLE, T. Ant Colony Optimization. Londres: The MIT Press, 2004.

DUDA, R. O. Pattern Classification and Scene Analysis (Wiley, New York, 2003).

FISHER, R. A. Iris Data Set. 1936, Disponível em: <http://archive.ics.uci.edu/ml/datasets/Iris>. Acesso em: 22 jun. 2007.

FREI, F. Introdução à análise de agrupamentos. São Paulo: Unesp, 2006.

FUKUNAGA, K. Introduction to Statistical Pattern Recognition (Academic Press, San Diego, 1990).

HAMMER, P. Distance-Based Classification Methods, Rutcor Research Report, disponível em http://rutcor.rutgers.edu/rrr.

HOSHEN, J and KOPELMAN, R. Phy. Rev. B, 14, 3438-3445 (1976).

JAIN, A. K. DUBES, R. C. Algorithms for Clustering Data. New Jersey: Prentice Hall, 1988.

LANDAU, D. P. A Guide to Monte Carlo Simulations in Statistical Physics, 2 ed., (Cambridge university press, New York, 2005).

MANTEGNA, R. N. Phys. A, 287, 412 (2000).

OLIVEIRA, M. J. Dinâmica Estocástica e Irreversibilidade, EDUSP, São Paulo, 2001.

OLIVEIRA, M. J. Termodinâmica, Livraria de Física, São Paulo, 2005.

METROPOLIS, N.; ULAM, S. The Monte Carlo Method. Journal of the American Statistical Association 44, p. 335-341, 1949.

RUSPINI, E. H. Numerical methods for fuzzy clustering. Information Sciences 2, p. 319-350, 1970.

SALINAS, S. R. A. Introdução à Física Estatística, 2 ed., EDUSP, São Paulo, (2005).
SOBOL, I. M. A primer for the Monte Carlo method, (CRC Press, Florida, 1994).

STAUFFER, D. Introduction of the Theory Percolation, 2 ed., Francis & Taylor Ltd, London, (2003).

UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/index.html, acesso em 13/01/2008.

ULAM, S. Neumann, J. V. and Monte Carlo Method, Los Alamos Science Special Issue 1987.