MÉTODO DE AGRUPAMENTO SUPERPARAMAGNÉTICO NO ENSEMBLE MICROCANÔNICO

Evert Elvis Batista de Almeida

Programa de Pós-Graduação em Biometria e Estatística Aplicada, UFRPE,

R. Dom Manoel de Medeiros, s/n - Dois Irmãos 52171-900 - Recife/PE

evertelvis@yahoo.com.br

Adauto José Ferreira de Souza

Depto. de Física - UFRPE ,

R. Dom Manoel de Medeiros, s/n - Dois Irmãos 52171-900 - Recife/PE

adauto@ufrpe.br

Resumo: *Muitas vezes um problema de otimização pode ser resolvido observando-se o comportamento de um sistema natural. Por exemplo, quando um sistema termodinâmico é resfriado, suas partículas (em geral, interagentes) se organizam numa estrutura que corresponde a um mínimo de energia. O método de agrupamento superparamagnético (SPC) proposto em 1996 por Domany[4] é baseado em propriedades do modelo físico-estatístico conhecido como modelo de Potts. Originalmente, o sistema de spins foi simulado por um método Monte Carlo no ensemble canônico, isto é, simulado em uma dinâmica estocástica (banho térmico) que à medida que a temperatura é reduzida ocorrem às transições de fase superparamagnética. Nas transições dão-se os agrupamentos de partículas similares, apresentando flutuações magnéticas medidas através de susceptibilidade. Associando as partículas (spins) de um ferromagneto granular aos itens de dados e uma correspondente função custo ao hamiltoniano do sistema. O sistema de spins é então "resolvido", e as configurações de equilíbrio termodinâmico, em uma dada temperatura, estão associadas às partições do conjunto de dados em grupos. Neste trabalho aplicamos uma extensão do método de agrupamento supermagnético no reconhecimento de padrões com a vantagem de eliminar os ruídos existentes nas massas de dados através dos modelos físicos aplicados no algoritmo.*

Palavras-chave: Clusterização, Monte Carlo, reconhecimento de padrões, Física estatística.

Introdução

As técnicas de análise de dados tomaram grande importância em diversas áreas da ciência devido às grandes quantidades de dados a ser discutido em certos experimentos, sendo necessário para o manipulador separar os termos com igual similaridade para não serem comparados com os distintos. Ao longo do tempo surgiram inúmeras técnicas e estas com suas peculiaridades e aplicações específicas. O presente trabalho visa apresentar um algoritmo de agrupamento de dados que teve suas origens na física estatística e se utilizando de um fenômeno físico chamado superparamagnetismo. Muitas das vezes alguns problemas de otimização podem ser resolvidos observando problemas semelhantes existentes na natureza. Este novo algoritmo fará o reconhecimento sobre um banco de dados[2] com 20000 itens que representam as 26 letras do alfabeto, sendo estas letras escritas em diferentes fontes e distorcidas em relação aos eixos. Este trabalho apresenta um forte apelo no reconhecimento de caracteres podendo ser aplicado futuramente em teste grafotécnicos.

Material e métodos

As substâncias ferromagnéticas são caracterizadas por possuírem uma magnetização (espontânea) que pode pesistir mesmo na ausência de um campo magnético. Esse comportamento é bem diferente daquilo que ocorre numa substância paramagnética em que a magnetizaçã só aparece quando se um aplica um campo magnético. Na substância paramagnética, a magnetização desaparece quando o campo se anula como vemos na figura [1].

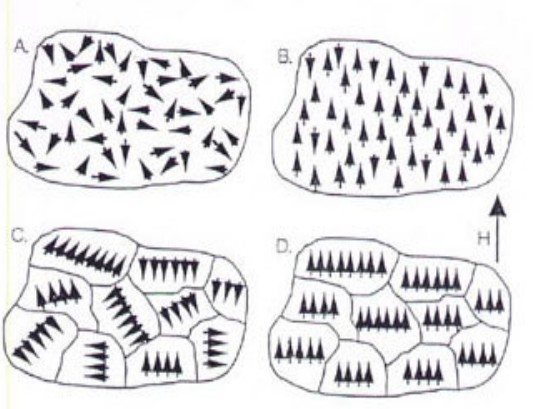

Figura1.Em A vemos a representação de material paramagnético, em B temos um material ferromagnético em C e D um material superparamagnético.

Se uma substância ferromagnética for aquecida a temperatura suficientemente alta ela perderá a magnetização espontânea e se comportará como uma substância paramagnética. Ou seja, ocorre uma transição de fase ferromagnética para a fase paramagnética. O estado superparamagnético é caracterizado pela existência de regiões (grupos) ordenadas do ferromagneto, porém descorrelacionadas entre si. Assim, o material como um todo apresenta um comportamento paramagnético. Com a diminuição da temperatura as regiões ordenadas passam a se influenciar mutuamente e a crescer de tamanho. Finalmente, formando um estado completamente ordenado para temperaturas suficientemente baixas. No outro extremo, o aumento de temperatura provoca o desordenamento dos grãos. Para temperaturas altas o suficiente, o sistema é completamente desordenado. Em temperaturas intermediárias, ocorre o ordenamento de regiões do material indicado por transições de fases.

O algoritmo desenvolvido no agrupamento de dados faz uso do método Monte Carlo, sugerido preliminar por Ulam[6]. O método Monte Carlo consiste em simular variáveis aleatórias, e visa obter soluções aproximadas para problemas de difícil solução. Podemos visualizar o método a partir do seguinte exemplo:

Queremos calcular a área da circunferência ilustrada na figura 02.

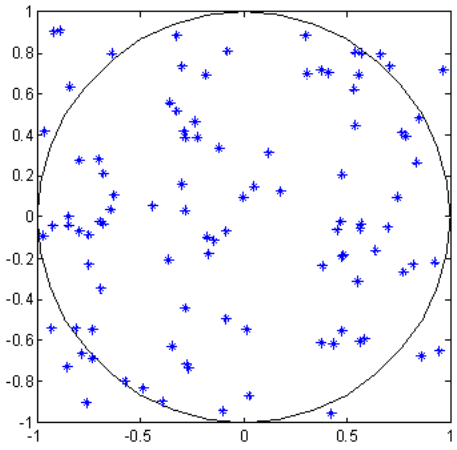

Figura 02: Mostra o sorteio de pontos aleatórios no uso do método Monte Carlo.

Confinamos a circunferência dentro de um quadrado, onde sorteamos pontos aleatórios tendo como limite os contornos do quadrado. Contamos o número de pontos que estão dentro do limite da circunferência e dividimos pelo total de pontos sorteados. A razão nos dá um resultado aproximado da área desejada. Este resultado tem seu erro reduzido com $\frac{1}{\sqrt{n}}$, onde n é o número de pontos sorteados, veja que para 100 pontos temos uma aproximação de 0.1, para 10000 pontos de 0.01 e assim por diante, este método para cálculos de áreas não é muito eficiente, mas para integrais de dimensões maiores não existe nenhum métodos que apresente vantagens na aplicação.

Em geral os itens de dados os quais faremos os agrupamentos, são representados em termos de vetores em um espaço multidimensional. Podemos estimar a similaridade entre dois itens utilizando uma distância que chamamos de Minkowski para os elementos x_i e x_k, k≠i, definida por:

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

Esta distância apresenta uma forma generalizada para tratar outras medidas de distâncias como a Euclidiana quando o λ=2, Manhattan para λ=1 e Chebyshev para λ=∞. Segundo Mingoti[5], a Minkowski λ≥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. Dada por uma hiper-esfera, possui a propriedade de dar maior ênfase a maior diferença entre uma única variável.

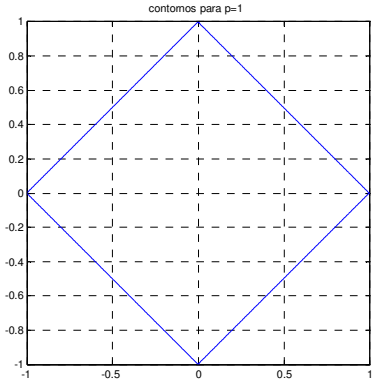

Figura 03:Mostra o contorno gerado pela distância Minkowski com λ=1.

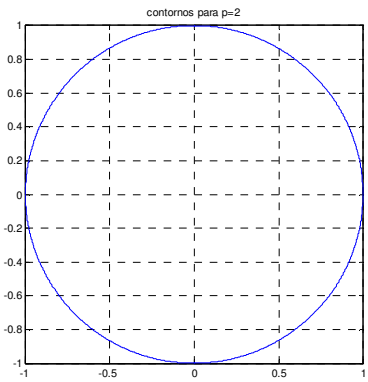

Figura 04:Mostra o contorno gerado pela distância Minkowski com λ=2.

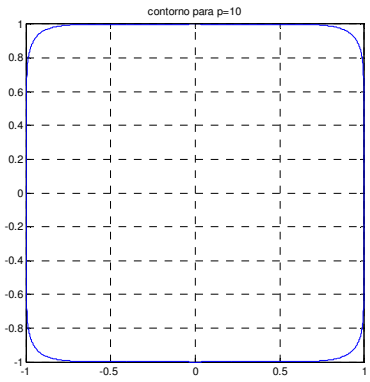

Figura 05:Mostra o contorno gerado pela distância Minkowski com λ=10.

Para o entendimento do método de clusterização superparamagnética no ensemble microcanônico, cabe uma breve descrição dos ensembles estatísticos. No ensemble microcanônico temos um sistema onde as partículas estão confinadas com valores de energia, volume e o número de moléculas são mantidos constantes. As partículas são idênticas e cada uma pode está em um determinado estado, que denominamos de orbital, r com energia ε(r). Os orbitais relativos a i-ésima partícula é denotada por r_i de modo que o estado total do sistema é definido pelo vetor (r_1, r_2,...,r_N). A energia total é simplesmente a soma das energias de cada partícula como vemos na expressão:

$$E = \sum_{i=1}^{N} \varepsilon(r_i)$$

A descrição probabilística do sistema se faz por meio da distribuição de probabilidade $P(r1, r2, \ldots, r_N)$ de encontrar no estado (r_1, r_2,...,r_N). vamos supor que a energia total do sistema é dada por U, nesta caso podemos dizer que só os estados em que a soma de energia seja igual a U podem ocorrer com a mesma probabilidade, conforme a expressão abaixo:

$$P(r_1, r_2,\ldots,r_N) = \frac{1}{W}$$

Em palavras, W é o numero de estados acessíveis em que a soma da energia é igual a U, W é dado por:

$$W = \sum_{n_1, n_2, \ldots, n_N}{}' \frac{1}{n_1!\, n_2! \ldots n_N}$$

O método cluterização superparamagnética no ensemble microcanônico associa cada ponto no espaço a uma variável de Potts S_i=1,2,...q. Os grupos são classificados de modo que quando um ponto é semelhante a outro eles são etiquetados, e quando são diferentes recebem uma penalidade. Desta forma introduzimos a função custo:

$$H(\{S\}) = \sum_{<i,j>} J_{ij}(1 - \delta_{s_i,s_j}), \qquad onde\ J_{ij} = \frac{1}{k}\exp\left(-\frac{d_{ij}^2}{2a}\right)$$

Onde $J_{i,j}$ é a constante de troca, $\tilde{k}$ é o número médio de vizinhos, d_{ij} a distância euclidiana entre os spins localizados nos sítios i e j, **a** é a distância média entre os pontos. O sistema então evolui de acordo com uma dinâmica estocástica onde calculamos as grandezas relevantes. Por exemplo, o comportamento da magnetização em função da temperatura pode indicar onde surge a ruptura no sistema, indicando a localização de uma transição de fase. Entre duas transições, temos uma fase relativamente estável e podemos identificar os grupos de itens similares. Desenvolvemos um programa computacional para aplicar o método, e introduzimos no programa uma nanomáquina que tenta alterar o estado de spins. Caso a mudança proposta diminua a energia, o demônio absorve a energia liberada pelo sistema. Por outro lado, se a energia do sistema aumenta com a mudança de estado, o demônio supre a energia necessária. No caso do demônio não ter energia suficiente, o novo estado é rejeitado. Dessa forma, a energia do sistema demônio+spins é mantida constante.

Resultados e discussões

Este método foi empregado para reconhecimento de letras, utilizamos um banco de dados desenvolvido pela marinha dos Estados Unidos, na figura 06 vemos as fontes utilizadas de cada item foram obtidos 16 atributos. Como a quantidade de itens do banco de dados original era muito grande era muito grande fizemos o teste apenas com as letras A, B e C, em um total de 2291 dados. Na figura 07 mostramos o crescimento da energia do sistema em função de temperatura. Na figura 08 observamos o gráfico do calor especifico em função da temperatura, temos a existência de vários picos ao longo do intervalo, isto denota que teremos vários agrupamentos, devido aos diversos tipos de fontes utilizados.

Conclusões

O método implementado mostrou-se bastante promissor nos primeiros testes realizado. Alguns ajustes no programa precisam ser feitos para obtermos uma melhor eficiência na classificação de dados. No reconhecimento de letras os resultados encontrados podem motivar outros trabalhos aplicados em testes grafológicos.

Figura06. Fontes utilizadas para a obtenção dos atributos.

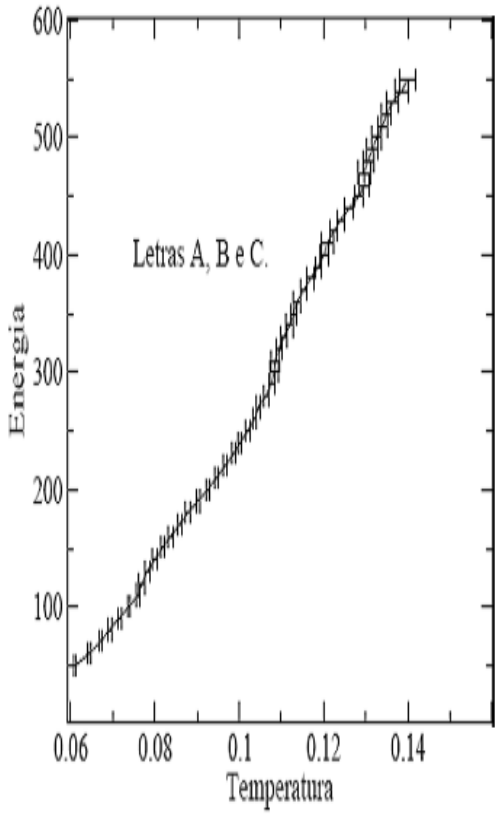

Figura 07: Mostra o aumento da energia em função da temperatura.

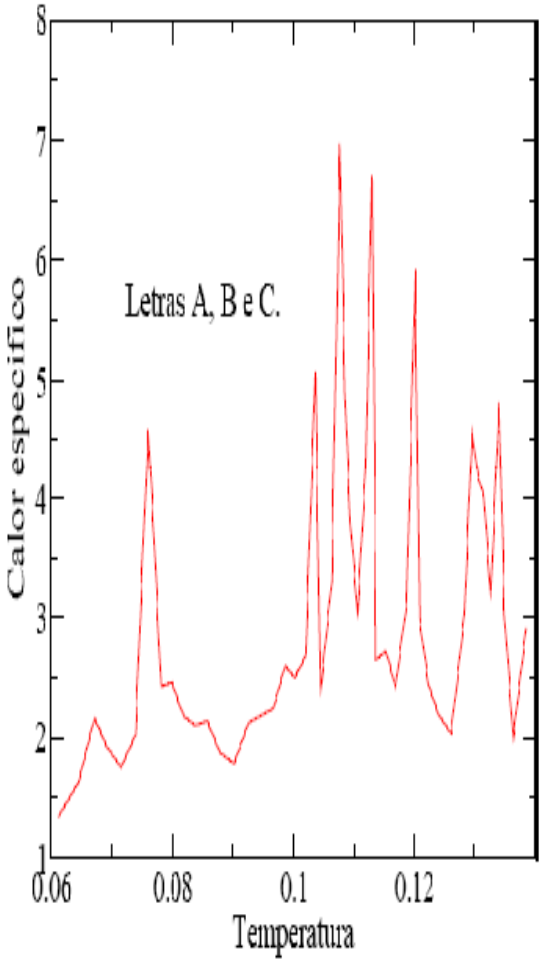

Figura 08:Mostra as transições de fase existente no sistema estudado.

Referências

[1]Creutz, M. 1983. Microcanonical Monte Carlo Simulation. Physical Review Letter.vol.50,19,1411-1414.

[2]Databases: www.mlearn.ics.uci.edu last seem on Oct. 17, 2007.

[4]E. Domany, M. Blatt and S. Wiseman, *Super-paramagnetic clustering of data,* Physical Review Letters 76, 3251 (1996).

[5]Mingoti, S. A., Análise de dados através de métodos de estatística multivariada. Belo Horizonte: Editora UFMG, 2007.

[6] S. Ulam, J. V. Neumann, and Monte Carlo Method, *Los Alamos Science* Special Issue *1987*.