MCAC - Monte Carlo Ant Colony: Um Novo Método de Data Clustering

José Domingos Albuquerque Aguiar
Programa de Pos-Graduação em Biometria, UFRPE,
52171-900, Recife, PE
domingos.aguiar@ig.com.br

Adauto José Ferreira de Souza
Depto. de Física, UFRPE
52171-900, Recife, PE
adauto@df.ufrpe.br

Resumo: Neste trabalho introduzimos um novo método para agrupamento de dados baseado no método de Monte Carlo [7] e na heurística da Ant Colony Optimization [2]. Aplicamos o método a dois conjuntos de dados reais e a um conjunto artificial de dados. Nossos resultados indicam que o esquema proposto é eficaz, robusto e relativamente simples para a indentificação de grupos em massas de dados.
Palavras-chave: Monte Carlo, Ant Colony e Data Clustering.

Introdução

Agrupamento é o processo de separar dados em classes, onde todos, ou a maior parte dos elementos pertencentes a mesma classe, tem um alto grau de similaridade. Normalmente o atributo utilizado para avaliar a semelhança é a distância entre pontos, ou seja, quanto maior a proximidade entre os pontos maior será as semelhanças entre eles.

Agrupamento ou data clustering [4] tem vasta gama de aplicação. Na biologia, data clustering pode separar espécies distintas de animais ou de plantas. Na química, pode separar substâncias diferentes baseado em características tais quais concentração, pH, temperatura e assim por diante. Na computação, pode ser aplicada na análise de imagens. Na economia, pode ser utilizada para identificar grupos de pessoas de baixa ou alta renda com baixo ou alto consumo, para fornecimento de crédito pessoal.

Material e Métodos

Ant Colony Optimization (ACO), é um processo de otimização estocástico inspirado no comportamento social de certa espécie de formiga [1]. Durante a procura por comida, as formigas empreendem caminhadas aleatórias a partir do ninho, ao encontrar uma fonte de alimento elas retornam ao formigueiro, percorrendo a trajetória em sentido inverso. As formigas marcam quimicamente o caminho percorrido com uma substância, secretada por elas, denominada de feromônio. Nas viagens seguintes, entre o ninho e a fonte de alimentos, as formigas reforçam a intensidade de feromônios nas trilhas. Porém, uma dada formiga ao deixar o formigueiro tende a escolher a trilha com maior intensidade de feromônio. Como o caminho mais curto entre a comida e o ninho foi o que recebeu maior quantidade de feromônio (a formiga que por ventura escolheu o menor caminho, foi e voltou antes das outras), os menores percursos têm maiores probabilidades de serem selecionados. Este efeito de realimentação faz com que as formigas utilizem o caminho mínimo entre a fonte de alimentos e o formigueiro.

O que foi dito acima torna-se mais claro com a descrição da seguinte experiência: montou-se um ninho de formigas e uma fonte de alimentos dentro de um aquário. Dois caminhos foram construídos entre o ninho e a fonte de alimentos, um mais curto que o outro. Inicialmente, as formigas escolhem um dos dois caminhos com igual probabilidade. Assim, aproximadamente metade das formigas segue o caminho mais longo e a outra metade segue o caminho mais curto (ver figura 1a). As formigas andam aproximadamente a uma velocidade constante. Logo, aquelas que escolheram o caminho mais curto realizam o percurso em menor tempo que as outras. O que resulta numa maior quantidade de feromônios depositada no caminho mínimo. Nas viagens subsequentes, a maioria das formigas escolherá o caminho mais curto, devido à alta concentração de feromônios nesta trilha (ver figura 1b).

O conjunto de dados é especificado por N vetores. Cada vetor pode ser visto como um ponto de um espaço euclidiano, cuja dimensionalidade iguala-se

ao número de características que descrevem um particular ítem do conjunto. Quanto menor a distância entre dois pontos desse espaço, maior a similaridade entre os ítens associados aos pontos.

A nossa proposta consiste em construir trilhas de feromônios conectando os pontos de dados com o requisito a seguir: pontos que pertencem ao mesmo grupo devem estar conectados por rastros mais intensos que os rastros entre pontos de grupos distintos. Para tanto, limitamos o comprimento máximo do percurso que uma dada formiga pode realizar, em cada ciclo do procedimento. Tomamos o comprimento máximo do percurso exponencialmente distribuído, de maneira que percursos muito longos são exponencialmente suprimidos. O nível de feromônios será, posteriormente, utilizado para idenficar os grupos, ou seja, caso a intensidade de feromônios em uma trilha que conecta dois pontos esteja acima de um determinado limiar, consideramos os dois pontos conectados ou pertencentes ao mesmo grupo. Caso contrário, a trilha é apagada e os dois pontos considerados pertencem a grupos distintos.

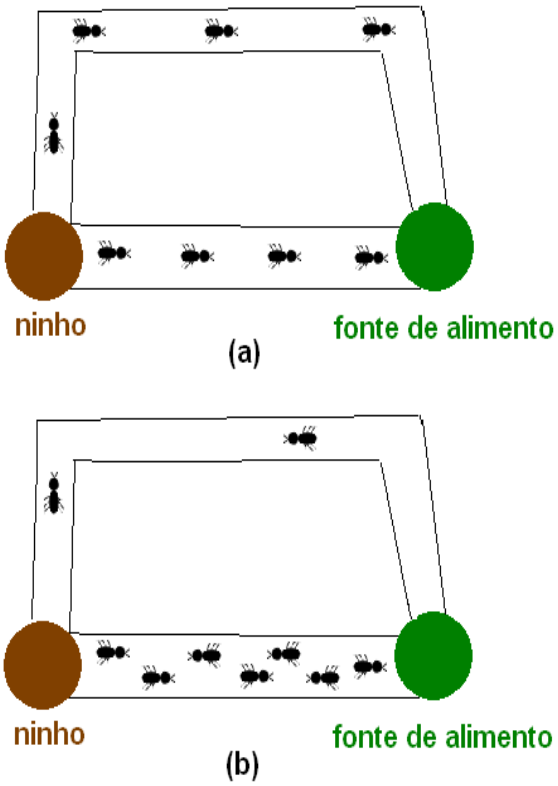

Figura l: Experiência realizada para observação do comportamento social das formigas: (a) Situação inicial onde as formigas escolhem um dos dois caminhos com igual probabilidade. (b) Segundo momento onde as formigas escolhem o caminho mais curto com probabilidade maior devido a alta concentração de feromônios.

Nossa proposta está esquematizada na figura 2, que consiste de quatro procedimentos básicos: Inicialização, Livre Exploração, Simulação e Finalização.

O procedimento de Inicialização tem cinco funções básicas:

1ª) Leitura das coordenadas dos N pontos.

2ª) Cálculo das distâncias geométricas entre todos os pontos, que resulta numa matriz N x N. Não são permitido pontos coincidentes na massa de dados para evitar divisão por zero na rotina de atualização da intensidade da trilha.

3ª) Faz o valor inicial de todas as trilhas iguais a zero;

4ª) Faz o número de formigas igual a N/3;

5ª) Cada formiga é colocada aleatoriamente em um dos pontos.

Figura 2: Seqüência do algoritmo.

Durante o procedimento de livre exploração, cada formiga visita aleatoriamente (distribuição uniforme) um terço dos pontos. Os pontos já visitados não poderão ser novamente visitados pela mesma formiga. Esta breve exploração aleatória fornece a escala de comprimento inerente à massa de dados. Esta informação será utilizada na etapa seguinte do algorítmo.

No procedimento de Simulação as trilhas de feromônios são construídas. Isto é, cada formiga visita um número de pontos que depende do comprimento máximo de seu percurso e atualiza a intensidade de feromônios. O número de simulações é um parâmetro livre. Cada formiga é livre para visitar de zero até N-1 pontos distintos aleatoriamente, em cada simulação. Como dito anteriormente, a cada ciclo, o comprimento máximo do percurso de cada formiga é determinado a partir de uma distribuição exponencial. A formiga só visitará o próximo ponto, se cumprir dois pré-requisitos: i) o número de pontos já visitados, incluindo o ponto inicial, deve ser menor que o total de

pontos; ii) o comprimento do caminho já percorrido pela formiga somado à distância que ela pretende percorrer, deve ser menor que o coprimento máximo do percurso selecionado para aquele ciclo. Caso o novo ponto passe pelos dois pré-requisitos, este ponto é visitado e o nível de feromônio da trilha é atualizado somando-se ao valor anterior, o quociente entre uma constante Q e a distância entre o ponto anterior e o novo ponto.

No procedimento de Finalização o valor médio das intensidades de feromônios nas trilhas é calculado. Nesse trabalho, tomamos como limiar de ativação de uma conexão entre dois pontos, o nível médio de feromônios vezes um certo fator, alfa. Todas as trilhas são analisadas, e aquelas que possuem um nível de feromônios menor que o limiar são consideradas não visíveis ou apagadas. Enquanto que, aquelas trilhas que possuem um nível de feromônios maior ou igual que o limiar são consideradas visíveis. Todos os pontos ligados por trilhas visíveis estão no mesmo grupo. Pequenos grupos, ou seja, grupos menores que 10% do total de pontos, são unidos a grupos maiores e mais próximos. Esse valor de 10% pode ser adaptado para cada caso.

Resultados e Discussão

O procedimento descrito acima foi aplicado a três conjuntos de dados, sendo um artificial e dois reais.

O primeiro conjunto de dados é uma base conhecida como Ruspini [6] (figura 3) que foi gerada artificialmente para testes de agrupamento. Ela é composta de 75 pontos bidimensionais, ou seja, cada ponto é composto de duas variáveis.

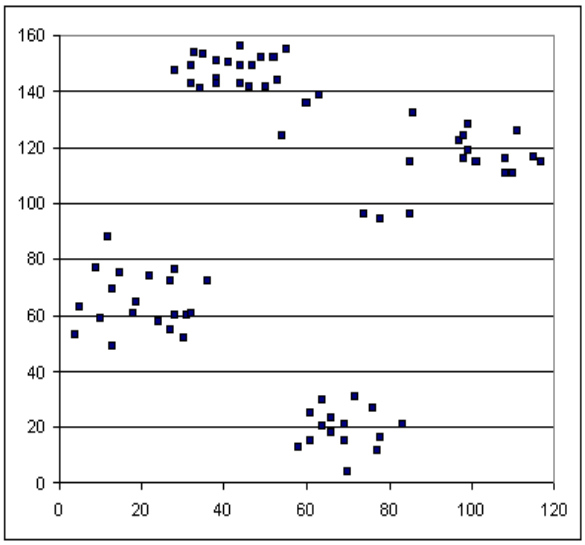

Figura 3: Base de dados Ruspini

Originalmente a base de dados Ruspini tem quatro classes (ou grupos) como a tabela 1 descreve.

Os dados do Ruspini foram analisados de acordo com nosso procedimento, com o número de simulações igual a 500. Na figura 4 temos uma imagem extraída do MCAC aplicado ao Ruspini com o limiar igual à média de todas as trilhas, ou seja, as trilhas com nível de feromônio abaixo da média são apagadas e as outras trilhas são visíveis. Nota-se claramente que com o limiar de 100% da média das trilhas já se identifica dois grupos distintos.

Tabela l: Classes da base de dados Ruspini.

	Nrº de pontos	Pontos
Classe 1	20	Do 1º ao 20º
Classe 2	23	Do 21º ao 43º
Classe 3	17	Do 44º ao 60º
Classe 4	15	Do 61º ao 75º

Aumentando o valor de 100% para 125%, temos como resultado apresentado na figura 5. O MCAC detectou quatro grupos distintos com 100% de acerto.

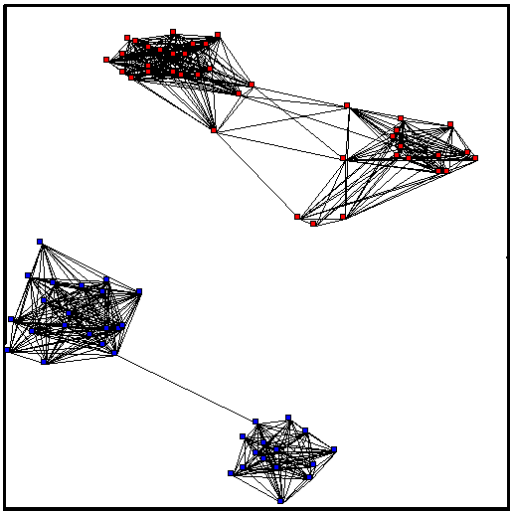

Figura 4: Ruspini com nível de trilas igual à média.

O segundo conjunto de dados é real e composta de dados faciais de 59 gorilas [5], entre machos e fêmeas (ver tabela 2). Diferentemente dos dados anteriores que possuíam apenas duas variáveis, esse banco de dados de gorilas é composto de treze variáveis.

Tabela 2: Classes da base de dados de gorilas

Classe	Nrº de pontos	Pontos
Machos	29	Do 1º ao 29º
Fêmeas	30	Do 30º ao 59º

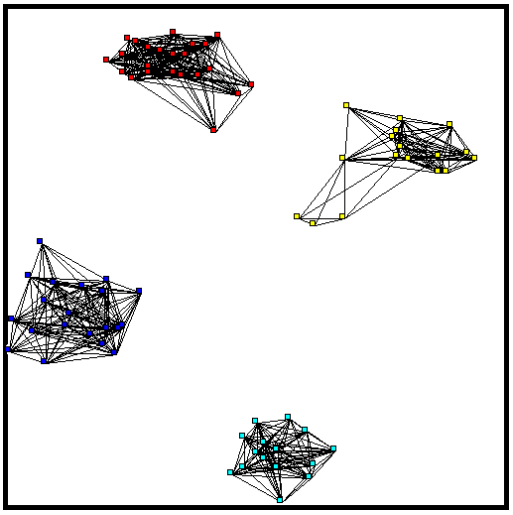

Figura 5: Ruspini com nível de trilhas igual a 125% da média.

Na figura 6, esta o gráfico dos 59 gorilas, com a utilização de duas das treze variáveis.

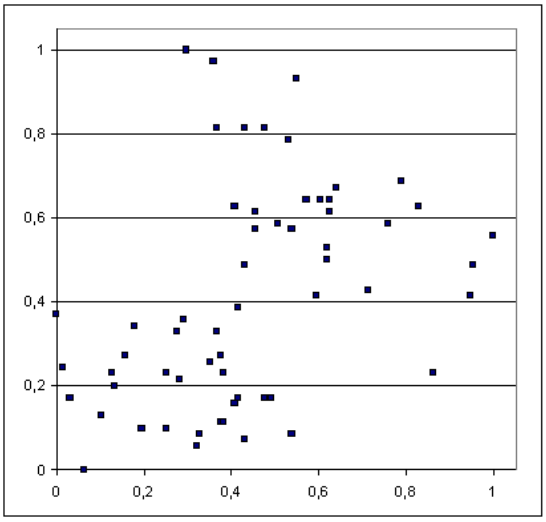

Figura 6: Gráfico com duas das treze variáveis de gorilas.

Os dados de gorilas foram rodados no MCAC também com 500 simulações. Na figura 7, temos uma imagem extraída do MCAC com nível de trilha igual a 155% da média das trilhas. Observam-se dois grupos onde o MCAC reconheceu os machos e as fêmeas com acerto de 100%.

O terceiro conjunto de dados é conhecido como Iris data que também é um conjunto de dados real, que esta disponível no repositório de dados da Universidade da Califórnia em Irvane [3]. Esse conjunto é composto de dados de 150 flores Iris, com três grupos distintos e com quatro variáveis. Mas, dois pontos do terceiro grupo (da Virginica) possuem as mesmas coordenadas (58, 27, 51, 19), sendo considerados assim o mesmo ponto. Desta forma para fins de nosso trabalho esse banco de dados é formado por 149 Iris de acordo com a tabela 3.

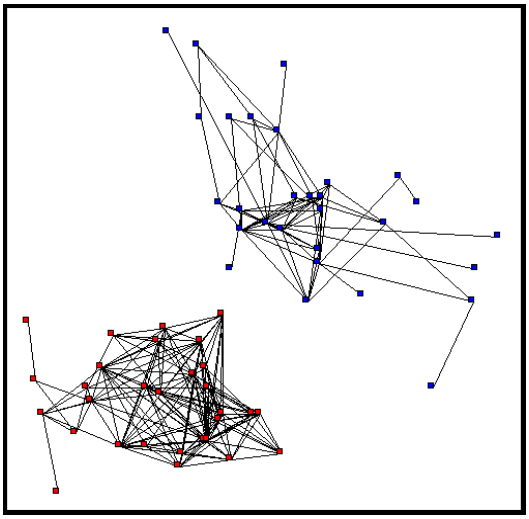

Figura 7: Dados de gorilas com nível de trilhas igual a 155% da média.

Tabela 3: Classes da base de dados Iris

Iris	Nrº de pontos	Pontos
Setosa	50	Da 1ª à 50ª
Versicolour	50	Da 51ª à 100ª
Virginica	49	Da 100ª à 149ª

Iris data é um problema de reconhecimento de espécies difícil e muito popular para testes de agrupamento. A figura 8 mostra o gráfico da Iris com três de suas quatro componentes (variáveis).

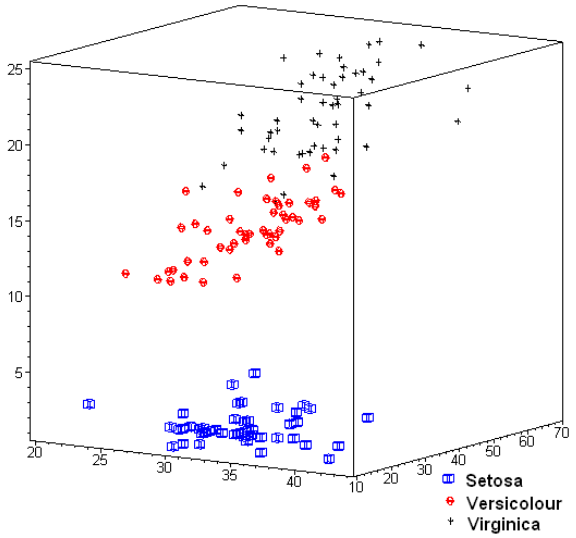

Figura 8: Gráfico das Iris com três de suas quatro componentes (variáveis).

Submetendo os dados das Iris à análise pelo MCAC com 500 simulações, obtivemos um percentual de acerto de 97,3%, classificando erroneamente apenas quatro das Iris, como mostra a tabela 4.

A figura 9 é uma imagem extraída do MCAC com duas dimensões e com o nível de trilha igual a 360% da média.

Tabela 4: Pontos mal classificados no Iris data.

Iris	Pontos mal classificados
Setosa	Nenhum
Versicolour	O ponto 107
Virginica	Os pontos 71, 73 e 84

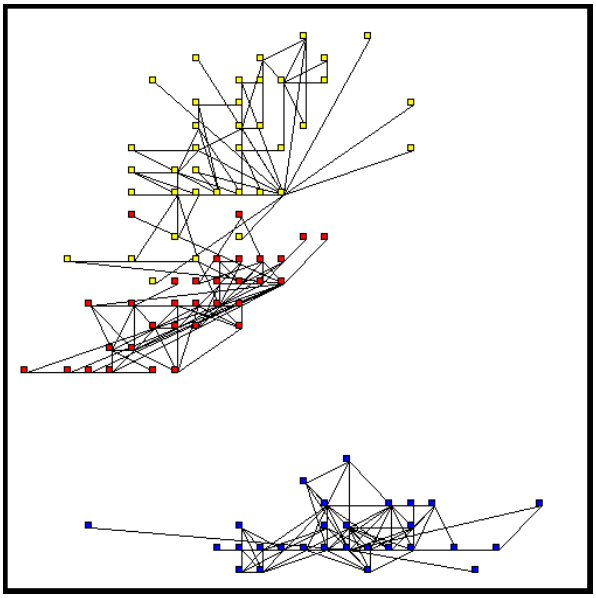

Figura 9: Dados das Iris com nível de trilhas igual a 360% da média

Conclusões

Nossos resultados indicam que o MCAC é um método eficiente para realizar a tarefa de agrupamento de dados. Além disso, destacamos algumas características do método:

1ª) É rápido;
2ª) Trabalha com várias dimensões;
3ª) Não precisa ter conhecimento do número de grupos com antecedência;
4ª) Não tem problema com clusters diferentes, onde número de elementos de um cluster seja várias vezes maior que o outro.

Atualmente, estamos trabalhando em um teste estatístico para verificar qual o nível ideal das trilhas em relação ao número de grupos.

Referências

[1] E. Bonabeau, M. Dorigo e G. Theraulaz, Inspiration for optimization from social insect behaviour**.** *Nature*, Vol. 406, 2006, pp. 39-42.

[2] M. Dorigo, Vittorio Maniezzo e Alberto Colorni, The Ant System: Optimization by a colony of cooperating agents, *IEEE Transactions on Systems, Man, and Cybernectics-Part B*, Vol. 26, No. 1, 1996, pp.1-13.

[3] R.A. Fisher, Iris Plants Database, *UCI Machine Learning Repository*, Irvine, 1988.

[4] A.K. Jain, M.N. Murty e P.J. Flynn, Data Clustering: A Review, *ACM Computing Surveys*, Vol. 31, No. 3, 1999, pp. 264-323.

[5] P. O'Higgins e I. L. Dryden, Sexual dimorphism in hominoids: further studies of craniofacial shape differences in Pan, Gorilla, Pongo, *Journal of Human Evolution*, Vol. 24, pp.183-205.

[6] E. H. Ruspini, A new approach to clustering, *Inf. Control*, Vol. 15, 1969, pp. 22–32.

[7] M. Sobol, O Método de Monte Carlo, *MIR*, Moscou, 1972.