SPOLM 2008

ISSN 2175-6295 Rio de Janeiro- Brasil, 05 e 06 de agosto de 2008.

AGRUPAMENTO DE DADOS COM ALGORITMO DE COLÔNIA DE FORMIGAS

Dilson Godoi Espenchitt
CASNAV/COPPE-UFRJ
Pr Brarão de Ladário S/N
espenchitt@casnav.mar.mil.br

Josir Cardoso Gomes
Fundação Getúlio Vargas - EBAPE/RJ
Praia de Botafogo, 190
josir@jsk.com.br

Nelson Ebecken
COPPE-UFRJ
Centro de Tecnologia - Bloco B - Sala 101 - Ilha do Fundão
Caixa Postal 68506 - CEP: 21945-970 RJ/RJ
nelson@ntt.ufrj.br

RESUMO

Este artigo mostra um algoritmo para agrupamento de dados, inspirado na Teoria de Colônia de Formigas, onde não precisamos selecionar a priori um número inicial de grupos.

O algoritmo foi implementado no software WEKA, desenvolvido em JAVA por uma equipe de pesquisadores da Universidade de Waikato (Nova Zelândia), com o código fonte disponível permitindo assim que novos algoritmos possam ser adicionados ao software.

Palavras-Chaves:Agrupamento de Dados, Colônia de Formigas, Software WEKA

Abstract

This paper shows a clustering data algorithm , supported by Ant Colony Theory, that doesn’t need a pre determined number of clusters.

This algorithm was implemented in WEKA software, developed in JAVA by a group of research from Waikato University, New Zeland, with code available allow new algorithm add to software .

Keywords: Cluster Analisys, Ant Colony, Software WEKA

1. INTRODUÇÃO

Na tarefa de agrupamento de dados, quando se necessita determinar o número de grupos em que uma base pode ser particionada, esta começa a se tornar bastante complexa, pois além de diversos parâmetros que se precisa ajustar, o que requer conhecimento específico do algoritmo que está sendo utilizado, alguns algoritmos requerem o número (n) de grupos (*clusters*) nos quais se quer particionar a massa de dados.

Após se aplicar diversos "n", utilizamos índices semi-empíricos tais como, Calinski e Harabasz, Critério Condorcet, *Cubic Clustering Criterion* e PBM, para ver qual o melhor "n". Este tipo de classificação não supervisionada encontra vasta literatura e tem sido tema de pesquisa ininterrupta. (MACHADO, 2002, PUNTAR, 2003, MORAES, 2004 , ANDRADE, 2004).

O presente artigo apresenta um algoritmo inspirado na teoria de Colônia de Formigas, onde não existe a necessidade de se utilizar um parâmetro "n" como entrada no algoritmo.

2. AGRUPAMENTO DE DADOS

O processo de agrupamento de objetos físicos ou abstratos em classes ou grupos de objetos similares é chamado de clusterização, que consiste na divisão dos dados em classes ou *cluster*s, de maneira que objetos dentro de um mesmo *cluster* tenham alta similaridade, mas objetos pertencentes a *clusters* diferentes sejam muito distintos (alta dissimilaridade). Em geral, para que os procedimentos possam medir a similaridade ou a diferença entre os objetos que estão sendo avaliados, diversas medidas são usadas. Estas medidas normalmente são baseadas em métricas de distância, e consideram adaptações para dados especiais (variáveis binárias, nominais, ordinais etc.), (Han & Clamber, 2006).

Mais especificamente, pode-se definir o problema de segmentação da seguinte maneira:

Seja X = { X_1, X_2,...., X_n} um conjunto com *n* elementos. Cada elemento do conjunto X, X_i é um vetor de R^p, representando um objeto através de *p* medidas que o descreve. Estas medidas são chamadas de atributos. Cada elemento de *X* deve ser considerado pertinente a um dos grupos C= {C_1, C_2,....,C_k}, aonde *k* é o número de *clusters*. Diz-se que *C* é a segmentação obtida para X, se podem ser observadas três características do conjunto *C*:

1. $C_1 \cup C_2 ..... \cup C_k = X$;
2. $C_i \neq \phi ; \forall_i ;$
3. $C_i \cap C_j = \phi ; \ \forall_{i=j} .$

A análise de agrupamento de dados é uma ferramenta bastante útil para o estudo e compreensão do comportamento de dados nas mais diferentes situações. Um exemplo disto é o caso de dados coletados através de pesquisas, onde pode-se obter uma quantidade extremamente grande de informações que, observados sob um contexto geral, podem não apresentar nenhum sentido, porém quando classificados e separados em grupos passam a fornecer informações com respeito ao comportamento de cada um destes grupos (HAIR *et al*,1998). Uma outra situação onde a análise de agrupamento de dados tem grande utilidade são os casos onde se deseja desenvolver hipóteses concernentes à natureza de uma massa de dados, ou mesmo examinar-se a veracidade de hipóteses previamente concebidas.

Dependendo dos interesses e objetivos que se deseje atingir, a visualização de agrupamento de dados nos possibilita obter uma preciosa ajuda para um rápido entendimento e assimilação de informações por meio das quais se possa avaliar fatores tais como: o quanto os grupos foram bem definidos. Como se diferenciam uns dos outros; seus tamanhos; a pertinência total ou parcial de uma amostra.

Uma análise de dados, portanto, pode ter basicamente dois objetivos primordiais, primeiramente poderá tratar-se de uma análise exploratória onde se buscará obter, dos dados em questão, as informações relevantes que eles possam conter e que não estão claramente mostradas em uma simples observação. Um outro objetivo seria uma análise confirmatória, onde os dados são utilizados para confirmar-se supostas informações previamente esperadas e que se acredita estarem neles contidas, e que podem ser confirmadas após uma maior exploração dos dados.

Agrupamento de dados é uma tarefa que procura segmentar populações heterogêneas em subgrupos ou segmentos homogêneos. Os registros são agrupados conforme alguma similaridade em si. (JAIN *et al*, 1999)

A simplicidade de uma estrutura se reflete em função do número de grupos, ou seja, uma estrutura é tão mais simples quanto menor possível for o número de grupos. Entretanto há que se considerar que a diminuição do número de grupos acarreta necessariamente uma diminuição também na homogeneidade dentro dos mesmos; portanto é necessário que exista um balanceamento entre o número de grupos e a similaridade entre eles. (JAIN *et al*, 1999).

Além disso, como parte do procedimento da análise de agrupamento de dados, pode-se executar uma redução dos mesmos, de forma a obtermos uma quantidade menor, porém extremamente objetiva das informações sobre uma população inteira de dados. A redução de dados nos fornece amostragem com informações a respeito de subgrupos menores, podendo-se assim obter uma descrição mais concisa e mais compreensível das observações, com uma perda mínima das mesmas.

Uma tarefa que pode ser associada ao agrupamento de dados é a identificação de pontos fora dos padrões. Os grupos representam registros ou objetos similares, entretanto existem muitos objetos que não apresentam uma forte pertinência a nenhum dos grupos em questão. Estes são exemplos de pontos fora dos padrões, que podem ser vistos como anomalias e pontos não bem assentados. (HAN & KAMBER, 2006).

Dependendo do segmento de negócios representado pelo conjunto de dados em análise, os pontos fora dos padrões podem, por exemplo, representar transações fraudulentas ou um comportamento não usual de cliente ou ainda uma tendência.

Pontos fora dos padrões são, portanto, as observações cujas características os identificam distintamente dos demais, ou seja, o ponto resultante da combinação de suas características apresenta-se comparativamente diferente dos demais. Pontos fora dos padrões não podem ser categoricamente caracterizados como benéficos tampouco como problemáticos, mas observados dentro do contexto da análise e avaliados em função do tipo de informação que poderão fornecer (HAIR *et al*,1998).

Um ponto fora do padrão poderá ser indicativo de alguma característica da população que não tenha sido revelada durante o curso normal da análise. Neste caso, ainda que distante da maioria das demais observações, apresenta um importante benefício na análise. Por outro lado, poderá também ser contrário aos objetivos da análise, distorcendo seriamente os testes estatísticos e, neste caso, trata-se de um ponto problemático, não sendo, portanto representativo da população. (HAIR *et al*,1998, HAN & KAMBER,2006)

Devido a estes diferentes aspectos na interpretação de um ponto fora do padrão, é de suma importância que se verifique os dados em análise a fim de se determinar os tipos de influências que os mesmos poderão causar.

Na evolução de um modelo de agrupamento de dados, o interesse primordial está concentrado nas seguintes questões(HAIR *et al*,1998):

- Como são os grupos similares entre si?
- A pertinência dos registros aos grupos mais prováveis é forte ou fraca?
- Existem registros de pontos fora dos padrões?
- Quais são as características típicas dos registros pertencentes a cada grupo? (Perfil dos grupos)
- O que diferencia cada grupo dos demais?

2.1. O PROBLEMA DA DEFINIÇÃO DO NÚMERO DE GRUPOS

Uma das grandes dificuldades nos problemas de agrupamento, como no método clássico *"K-Means"* certamente, é a exigência do valor de k, onúmero de grupos em que a massa de dados será dividida, como parâmetro inicial. Para uma aplicação real, em grandes bases de dados, esta informação normalmente é desconhecida, podendo ser necessária uma prévia análise de um especialista para que houvesse possibilidade de percepção da quantidade de grupos em que a coleção poderia dividir-se de forma satisfatória, o que por si só já representa uma grande carga de trabalho que, até mesmo, diante de uma base muito grande poderia tornar a análise inviável ou desinteressante.

Existem vários critérios para a determinação do número de grupos e quase todos funcionam da seguinte maneira: realizar o agrupamento dos dados considerando 2 grupos e calcular o valor de uma função proposta que tenha o número de grupos como um de seus parâmetros, realizar o agrupamento dos dados considerando 3 grupos e calcular o valor da mesma função, repetir este procedimento até atingir um número máximo de grupos estabelecido.

Há diversos índices semi-empíricos que podemos usar, tais como, Calinski e Harabasz, Critério Condorcet, *Cubic Clustering Criterion* e PBM, para ver qual o melhor "n", número de grupos. Este tipo de classificação não supervisionada encontra vasta literatura e tem sido tema de pesquisa ininterrupta. (ANDRADE, 2004, MACHADO, 2002, MORAES, 2004, PUNTAR, 2003).

O agrupamento que ocasionar o valor máximo (ou, em alguns casos, mínimo) da função, deve ser considerado como o melhor agrupamento possível para a base de dados.

3. ALGORITMO DE COLÔNIA DE FORMIGAS PARA AGRUPAMENTO DE DADOS.

A clusterização inspirada em Colônia de Formigas foi proposta inicialmente, por DENEUBOURG *et al*, 1991.

A principal vantagem deste algoritmo é o fato de que não é necessária nenhuma informação inicial a respeito da massa de dados que iremos particionar, e os parâmetros necessários para sua execução são o número de "formigas" que serão usadas e o número de ciclos que serão executados. Além das características descritas anteriormente, de diversos relatos na literatura do seu uso com sucesso para clusterização. : (AZZAG *et al*, 2003, FREITAS, 2001, HANDL, 2003, HANDL *et al*, 2003, HANDL & KNOWLES,$2004_{a,}$ 2004_{b}, HANDL *et al*, 2005, HANDL & MEYER,2007, MONMARCHÉ, 1999, PULIDO & COELLO, 2004).

A implementação de um algoritmo ACO para Agrupamento de Dados, segue as seguintes etapas propostas por (MONMARCHÉ, 1999):

(i) cada objeto é representado por um vetor de dimensão *n*, onde *n* representa seus atributos;

(ii) os objetos são distribuídos aleatoriamente em um "tabuleiro", dimensionado de acordo com o tamanho da nossa população;

(iii) durante a execução do algoritmo, os objetos podem ser "empilhados" na mesma célula, constituindo "pilhas"; deste modo, as pilhas representam um *cluster;*

(iv) a distância entre dois objetos é calculada como sendo a distância euclidiana entre dois vetores em R^n;

(v) o centro do *cluster* é determinado pelo centro de massa dos objetos que o compõem (centróide);

(vi) a distância de dois *clusters* é dada pela distância dos seus centróides;

(vii) um número prédeterminado de "formigas", move-se no tabuleiro a cada interação, podendo tomar diferentes ações: largar ou pegar um objeto de acordo com seu estado:

- se ela não carrega nenhum objeto ela pode: pegar um objeto na célula vizinha ou pegar o objeto com maior dissimilaridade (objeto que apresenta a maior distância ao centróide da pilha) da "pilha" vizinha;
- se ela carrega um objeto ela pode: soltar o objeto em uma célula vizinha vazia, soltar o objeto em uma célula ocupada por um objeto e formar uma

"pilha" ou soltar o objeto em uma célula ocupada por uma pilha, onde a colocação do objeto não interfira no deslocamento do centróide.

(viii) O procedimento vii é repetido até atingirmos um número de interações pré-estabelecidas.

3.1. IMPLEMENTAÇÃO DO ALGORITMO DE COLÔNIA DE FORMIGAS PARA AGRUPAMENTO DE DADOS

O Algoritmo de Colônia de Formigas para Agrupamento de Dados foi implementado como uma ferramenta do software WEKA.

A implementação foi feita utilizando a linguagem JAVA, aproveitando as classes e os objetos do WEKA e usando o software ECLIPSE, versão 3.1 e posteriormente a versão EUROPA, como plataforma de desenvolvimento.(ESPENCHITT,2008)

4. O *SOFTWARE* WEKA

O pacote Weka (Waikato Environment for Knowledge Analysis) é formando por um conjunto de implementações de algoritmos de diversas técnicas de Mineração de Dados, é uma ferramenta de KDD que contempla uma série de algoritmos de preparação de dados, de aprendizagem de máquina (mineração) e de validação de resultados. WEKA foi desenvolvido na Universidade de Waikato na Nova Zelândia,sendo escrito em Java e possuindo código aberto disponível na Web.

Por estar implementado na linguagem Java, que tem como principal característica ser portável. Desta forma, pode rodar nas mais variadas plataformas, aproveitando os benefícios de uma linguagem orientada a objetos como modularidade, polimorfismo, encapsulamento, reutilização de código dentre outros.

Grande parte de seus componentes de software são resultantes de teses e dissertações de grupos de pesquisa desta universidade. Inicialmente, o desenvolvimento do software visava a investigação de técnicas de aprendizagem de máquina, enquanto sua aplicação inicial foi direcionada para a agricultura, uma área chave na economia da Nova Zelândia.

O sistema possui uma interface gráfica amigável e seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado. Grande parte de seus recursos é acessível via sua GUI, sendo que os demais podem ser utilizados programaticamente através de API´s. Foi disponibilizada também uma abrangente

documentação online do código fonte. Por ser escrito em Java, o código pode ser rodado em diferentes plataformas, conferindo uma boa portabilidade ao software.

O WEKA possui um formato próprio de arquivo de dados, o ARFF, o qual descreve o domínio do atributo, pois o mesmo não pode ser obtido automaticamente pelo seu valor.

Antes de aplicar os dados a qualquer algoritmo do pacote WEKA, estes devem ser convertidos para o formato ARFF que consiste basicamente de duas partes. A primeira contém uma lista de todos os atributos, onde se define o tipo do atributo ou os valores que ele pode representar, quando se utiliza valores estes devem estar entre "{ }" separados por vírgulas. A segunda parte consiste das instâncias, ou seja, os registros a serem minerados com o valor dos atributos para cada instância separado por vírgula, a ausência de um item em um registro deve ser atribuída pelo símbolo "?".(WITTEN &FRANK, 2005)

Podem-se usar programas de planilhas eletrônicas e banco de dados os quais permitem exportar os dados em um arquivo onde as vírgulas são os separadores. Existem também ferramentas desenvolvidas que permitem exportar os dados de planilhas Microsoft Excel para o formato ARFF (Gomes et al, 2006)

Uma vez feito isso, é necessário apenas carregar o arquivo em um editor de texto e adicionar o nome do conjunto de dados usando @relation nome_do_conjuto_de_dados, para cada atributo usa @attribute o nome do atributo e o tipo do atributo (real, nominal, categórico, binário etc) e após colocar uma linha com @data e logo em seguida os dados em si, salvando o arquivo como texto puro com extensão ARFF.

5. ESTUDO DE CASO

Com o propósito de demonstrar a capacidade do algoritmo para se determinar o número de *clusters* foram realizados quatro experimentos empregando-se o algoritmo de Colônia de Formigas para Agrupamento de Dados implementado no *software* WEKA e feita a comparação com o algoritmo *Expectation Maximization* (EM), além de ser um algoritmo residente do WEKA e fornecer o número final de *clusters* sem a necessidade de se precisar fornecer um número inicial.

Foram avaliados dois pontos: o acerto do número de *clusters* e a acurácia(taxa de acerto dos elementos nos grupos) .

As bases de teste usadas foram: íris, soybean, bank e mushroom obtidas no software WEKA e no repositório de dados da UCI (Universidade de Califórnia Irvine), disponível em UCI, 2007.

A Tabela 1 mostra as características das bases : números de atributos e número de exemplos das bases de dados que foram usadas.

Tabela 1 - CARACTERÍSTICAS DAS BASES DE DADOS

Base	Atributos	Exemplos
Iris	4	150
Soybean	34	683
Bank	16	600
Mushroom	22	8124

Tabela 2 - RESULTADOS

Base	NºCluster	NºCluster Alg. Implementado	NºCluster EM	AcuráciaAlg. Implementado
Iris	3	3	5	90%
Soybean	19	19	4	90%
Bank	6	6	8	95%
Mushroom	2	2	13	90%

Nos casos analisados o Algoritmo de Colônia de Formigas para Agrupamento de Dados implementado apresentou a melhor taxa de acerto que o algoritmo EM.

No algoritmo EM não foi avaliada a acurácia, pois sua taxa de acerto do número de *clusters* foi nula, desta forma não faz sentido computar a acurácia.

6. CONCLUSÃO

O Algoritmo de Colônia de Formigas para Agrupamento de Dados implementado no software WEKA atende ao objetivo de se determinar o número de *clusters* de uma massa de dados em um número desconhecido de grupos, sem termos que selecionar a priori um número inicial de grupos.

Nos casos analisados, o Algoritmo de Colônia de Formigas para Agrupamento de Dados apresentou melhor taxa de acerto que o algoritmo EM.

Quanto à acurácia, o algoritmo EM não foi avaliado, pois a taxa de acerto do número de clusters foi nula, não fazendo sentido a comparação com o Algoritmo de Colônia de

Formigas para Agrupamento de Dados que sempre apresentou um resultado muito bom neste quesito.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, L. P., **Procedimento Interativo de Agrupamento de Dados.** Tese M. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.

AZZAG, H., GUINOT, C., VENTURINI,G., **How to use ants for hierarchical clustering.** Fourth international workshop on Ant Colony Optimization and Swarm Intelligence, p.350-357, LNCS 3172, Brussels, Belgium,2004.

DENEUBOURG, J.L.,GOSS, S., Franks N., SENDOVA, F. A., Detrain, C., Chr´etien L.,**The Dynamics of Collective Sorting Robot-Like Ants and Ant-LikeRobots**. From Animals to Animats: Proc. of the 1st Int. Conf. on Simulationof Adaptive Behaviour. 1990.

ESPENCHITT, D. G., **Segmentação de Dados em um Número Desconhecido de Grupos Utilizando Algoritmo de Colônia de Formigas.** Tese D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2008.

FREITAS, A. A., **A survey of evolutionary algorithms for data mining and knowledge discovery**. Advances in Evolutionary Computation, A. Ghosh & S. Tsutsui(eds.), Springer-Verlag, 2001.

GOMES, Josir Cardoso ; LEVY, Ariel ; LACHTERMACHER, Gerson . **Segmentação do Censo Educacional 2000: Utilizando Técnicas de Mineração de Dados**. In: Luiz Fernando de Paula; Léo da Rocha Ferreira; Milton de Assis. (Org.). Perspectivas Para a Economia Brasileira. 1 ed. Rio de Janeiro: Editora da Universidade do Estado do Rio de Janeiro, 2006, v. 1, p. 391-409.

HAN, J., KAMBER, M., **Data Mining Concepts and Techniques**. Morgan Kaufmann Publishers, San Francisco, USA, 2006.

HAIR, J.F. JR, A ., R. F., TATHAM, R. L., **Multivariate Data Analises**. 5 ed. Prentice Hall, Inc,USA,1998.

HANDL, J., **Ant-based methods for tasks of clustering and topographic mapping**: **extensions, analysis and comparison with alternative techniques**. Masters Thesis, Universität, Erlangen-Nürnberg, Erlangen, Germany,2003 .

HANDL,J., KNOWLES, J., DORIGO, M., **On the performance of ant-based clustering. Design and application of hybrid intelligent systems.** Frontiers in Artificial intelligence and Applications 104. Pages 204-213. 2003.

HANDL, J., KNOWLES, J., **Multiobjective clustering with automatic determination of the number of clusters.** Technical Report TR-COMPSYSBIO-2004-02. UMIST, Manchester, UK, 2004_a.

HANDL, J., KNOWLES, J., **Evolutionary multiobjective clustering**. Proceedings of the Eighth International Conference on Parallel Problem Solving from Nature (PPSN VIII). Pages 1081-1091. LNCS 3242. Springer-Verlag, 2004_b.

HANDL, J., KNOWLES, J., **Exploiting the trade-off -- the benefits of multiple objectives in data clustering.** Third International Conference on Evolutionary Multi-Criterion Optimization, 2005.

HANDL, J., KNOWLES, J., DORIGO, M**., Ant-based clustering and topographic mapping.** Artificial Life 11.2. ,2005**.**

HANDL J., MEYER B., **Ant-based and swarm-based clustering**.,. Proceedings of the IEEE Swarm Intelligence Symposium, SIS 2007.2007.

JAIN, A.K., MURTY, M.N., FLYNN P.J., **Data Clustering: A Review**, ACM Computing Surveys, Vol. 31, No. 3, September 1999

MONMARCHÉ**, N**., **On data clustering with artificial ants**. A.A. Freitas, editor, AAAI-99 & GECCO-99 Workshop on Data Mining with Evolutionary Algorithms: Research Directions, pages 23-26, Orlando, Florida, July 18,1999.

MACHADO FILHO, O. M., **Exploração e Análise de Agrupamento de Dados.** Tese M. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2002.

MORAES, D. R. S. **Inteligência Computacional na Classificação Litológica.** Tese M. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.

PUNTAR, S. G., **Métodos e Visualização de Agrupamento de Dados**, Tese M. Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2003.

UCI Machine Learning Repository-<http://ww.ics.uci.edu/~mlearn/>Acessado em: 20 dez 2007

Weka Software. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Disponível em : http://www.cs.waikato.ac.nz/ml/weka/.
Acesso em: jun/ 2006.

WITTEN, I.H., FRANK, E., Data **Mining: Practical machine learning tools and techniques with Java implementations**. San Francisco: Morgan Kaufmann Publishers. 2nd Edition, 2005.

.