MAFIA: *Merging of Adaptive Finite Intervals*

Elaine Ribeiro de Faria

Análise de Agrupamento de Dados

ICMC-USP – Dezembro 2010

Sumário

- Introdução
- Visão Geral e Objetivos do MAFIA
- Algoritmo *Grid* Adaptativo
- Algoritmo MAFIA
- Algoritmo pMAFIA
- Detalhes dos testes executados

Referências utilizadas

- Nagesh H., Goil S. e Choudhary A., **Adaptive Grids for Clustering Massive Data Sets**. In *SDM*, 2001.
- Nagesh H., Goil S., Choudhary A. e Choudhary A., **A Scalable Parallel Subspace Clustering Algorithm for Massive Data Sets**, Proceedings of the 2000 International Conference on Parallel Processing (ICP'00), 2000.
- Nagesh H., **High Performance Subspace Clustering for Massive Data Sets**, Master Thesis, Northwestern University, Evanston, 1999.
- Goil S., Nagesh, H. e Choudhary A., **MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets**, Technical Report CPDC-TR-9906-010, Center for Parallel and Distributed Computing, Department of Electrical & Computer Engineering, Northwestern University, June 1999.

Introdução

- Técnicas de agrupamento de dados são muito utilizadas em grandes bancos de dados com grande número de dimensões
- Essas técnicas devem tratar diversas questões
  - Escalabilidade com grandes bases e com alta dimensionalidade
  - Representação efetiva dos grupos
  - Uso de poucos parâmetros do usuário
  - Grupos podem estar embutidos em um subespaço do espaço de dados total

Visão geral do MAFIA

- Proposta baseada em densidade e *grid* para detectar grupos em subespaços
  - Baseado no algoritmo CLIQUE
  - Densidade → considera os grupos como regiões de alta densidade separados por regiões de baixa densidade
  - *Grid* →O espaço multidimensional é dividido em um grande número de regiões hiper-retangulares
    - regiões que tem mais pontos que um específico limiar são identificadas como densas
  - As regiões hiper-retangulares densas que são adjacentes a outras são unidas para encontrar os clusters

Objetivo do MAFIA

- Objetivo
  - Usar *grids* adaptativos para calcular os grupos em subespaços
    - *Grids* uniformes → muito esforço computacional e grupos de baixa qualidade
  - Usar um algoritmo *bottom-up* para agrupamento de subespaços
    - Calcula unidades densas em todas as dimensões
    - Combina-as para gerar unidades densas em dimensões maiores
    - Possui uma versão paralela

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i (A_i é o *i-ésimo* atributo)
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Termos usados

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
  Divida D_i em *janelas* de tamanho *x* (pequeno)

fim

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
Divida D_i em *janelas* de tamanho *x* (pequeno)

fim

Figura adaptada de Leung et al. (2005).

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
Divida D_i em *janelas* de tamanho *x* (pequeno)
Calcule o histograma para cada unidade de Ai, e atribua ao valor da *janela* o valor máximo encontrado na *janela*

Uma passagem sobre os dados

fim

Algoritmo – *Grid* Adaptativo

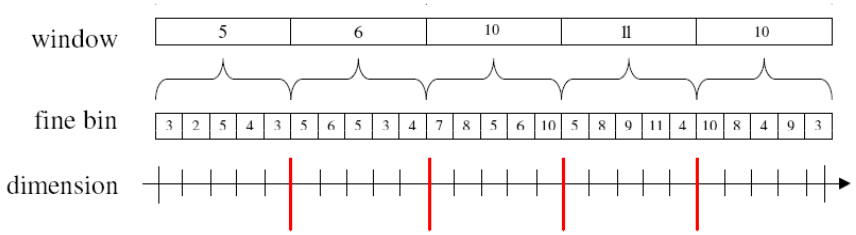


Figura adaptada de Leung et al. (2005).

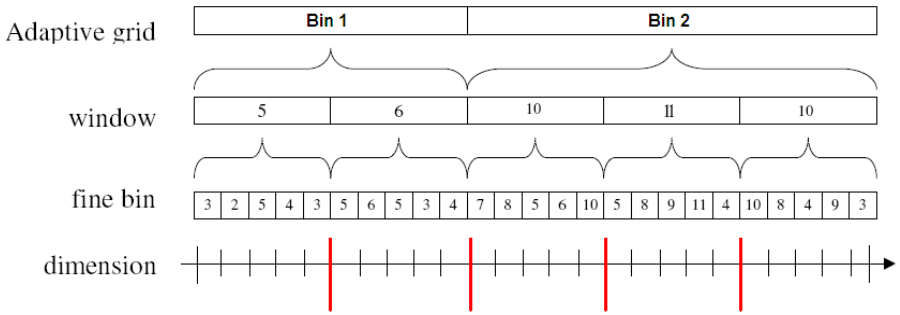
Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
Divida D_i em *janela* de tamanho *x* (pequeno)
Calcule o histograma para cada unidade de Ai, e atribua ao valor da *janela* o valor máximo encontrado na *janela*
Da esquerda para a direita, una duas unidades adjacentes se elas estão dentro de um limiar β

fim

Algoritmo – *Grid* Adaptativo



β utilizado = 20%

Figura adaptada de Leung et al. (2005).

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
  Divida D_i em *janelas* de tamanho *x* (pequeno)
  Calcule o histograma para cada unidade de Ai, e atribua ao valor da *janela* o valor máximo encontrado na *janela*
  Da esquerda para a direita, una duas unidades adjacentes se elas estão dentro de um limiar β
  /*Se o número de *bins* é 1 tem-se uma dimensão equi-distribuída*/
  Se (número de bins == 1)

fim

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
  Divida D_i em *janelas* de tamanho *x* (pequeno)
  Calcule o histograma para cada unidade de Ai, e atribua ao valor da *janela* o valor máximo encontrado na *janela*
  Da esquerda para a direita, una duas unidades adjacentes se elas estão dentro de um limiar β
  /*Se o número de bins é 1 tem-se uma dimensão equi-distribuída*/
  Se (número de bins == 1)
    Divida a dimensão Ai em um número fixo de partições iguais

fim

Algoritmo – *Grid* Adaptativo

D_i – Domínio de A_i
N – número total de registros na base de dados
a – tamanho de um *bin* genérico

Para cada dimensão A_i, i ∈ (1,..,d)
  Divida D_i em *janelas* de tamanho *x* (pequeno)
  Calcule o histograma para cada unidade de Ai, e atribua ao valor da *janela* o valor máximo encontrado na *janela*
  Da esquerda para a direita, una duas unidades adjacentes se elas estão dentro de um limiar β
  /*Se o número de bins é 1 tem-se uma dimensão equi-distribuída*/
  Se (número de bins == 1)
    Divida a dimensão Ai em um número fixo de partições iguais
  Calcule o limiar para cada *bin* de tamanho a como $\frac{\alpha a N}{D_i}$
fim

Algoritmo – *Grid* Adaptativo

$$\frac{\alpha\, a\, N}{D_i}$$

Tamanho do *bin*

Domínio do Atributo A_i

No exemplo anterior o primeiro *bin* encontrado tem tamanho 10

O Domínio D_i do *i-ésimo* atributo é de tamanho 25

Logo 10 / 25 = 0,4

Algoritmo – *Grid* Adaptativo

$$\frac{\alpha\, a\, N}{D_i}$$

Número de elementos da base

No exemplo anterior temos

$a/D_i = 0,4$ N = 147

0,4 * 147 = 58,8

Se o *bin* ocupa 40% do tamanho do domínio, espera-se encontrar 40% dos dados nesse *bin*

Algoritmo – *Grid* Adaptativo

Fator de dominância do grupo

$$\frac{\alpha\, a\, N}{D_i}$$

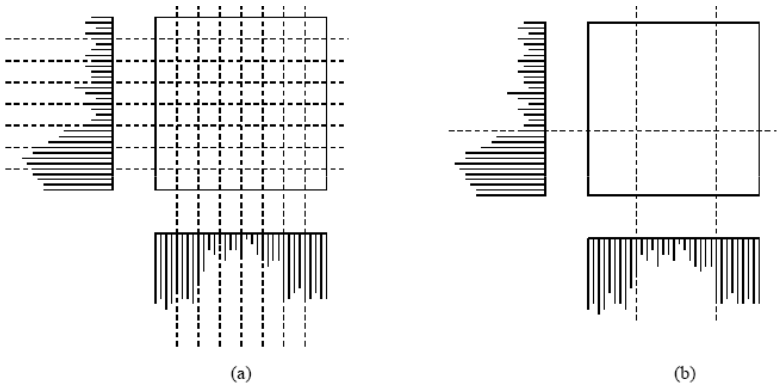
Um *bin* é denso se a fração do total de pontos contidos no *bin* é significantemente maior do que o valor esperado se os dados fossem uniformemente distribuídos no espaço de dados

Continuando o exemplo anterior

$a/Di * N = 58,8$ α = 1,5

58,8* 1,5 = 88,2

Efeitos do *grid* adaptativo



(a) Uniform grid size (b) Adaptive grid size

Figura extraída de Nagesh et al. (2000).

MAFIA

N – número de registros
d – dimensionalidade dos dados
A_i – *i-ésimo* atributo
B – número de registros que cabem na memória

Termos usados

MAFIA

N – número de registros
d – dimensionalidade dos dados
A_i – *i-ésimo* atributo
B – número de registros que cabem na memória

Leia os dados em blocos de B registros

MAFIA

N – número de registros
d – dimensionalidade dos dados
A_i – *i-ésimo* atributo
B – número de registros que cabem na memória

Leia os dados em blocos de B registros

Idade	Salário	Nível Superior	Anos trabalho	
20	2.000,00	0	1	Registro 1
30	4.000,00	1	6	Registro 2
40	3.000,00	0	15	Registro 3
...	....	....	....	
20	400,00	0	1	Registro B

MAFIA

N – número de registros
d – dimensionalidade dos dados
A_i – *i-ésimo* atributo
B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão A_i, $i \in (1,..,d)$
Determine os intervalos adaptativos usando o histograma em cada dimensão A_i, $i \in d$ e determine o nível do limiar

Algoritmo – *Grid* Adaptativo

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os *bins* encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1

As CDU's de 1 dimensão serão os *bins* encontrados em cada dimensão

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os *bins* encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (não forem encontradas mais unidades densas)

fim

Podem ser encontradas unidades densas k-dimensionais, onde k < d

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – i-ésimo atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os *bins* encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
  Se (k > 1) {Encontre-unidades-densas-candidatas();}

fim

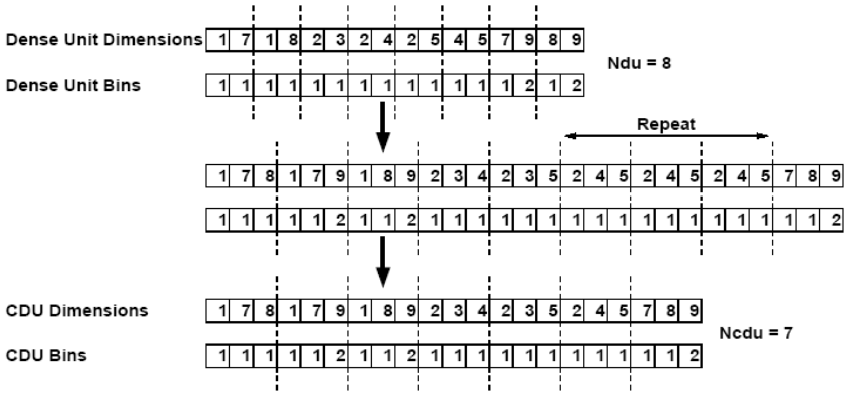
As unidades densas candidatas k-dimensionais, são formadas a partir das unidades densas (k-1) dimensionais

MAFIA

- CDU's em k-dimensões são obtidas unindo quaisquer duas células densas, representadas por um conjunto ordenado de (k-1) dimensões, tais que elas dividam qualquer uma das (k-2) dimensões
- Ex: $\{a_1, b_7, c_8\}$ e $\{b_7, c_8, d_9\}$

$$\{a_1, b_7, c_8, d_9\}$$

MAFIA



Dimensão corrente k = 3 Dimensão dos dados: 10

Ndu → número de unidades densas (2 dimensões)

Ncdu → número de unidades candidatas densas

Figura extraída de Nagesh et al. (2001)

MAFIA

N – número de registros d – dimensionalidade dos dados

A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os *bins* encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
  Se (k > 1) {Encontre-unidades-densas-candidatas();}
  Leia os dados em blocos de B registros e para cada registro popule as unidades candidatas densas
fim

Para cada valor de K uma varredura no banco de dados é necessária

MAFIA

N – número de registros d – dimensionalidade dos dados

A_i – i-ésimo atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os bins encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
  Se (K > 1) {Encontre-unidades-densas-candidatas();}
  Leia os dados em blocos de B registros e para cada registro popule as unidades candidatas densas
  Identifique-unidades-densas():
fim

MAFIA

- Uma CDU é densa se a contagem do seu histograma é maior que o limiar de todos os *bins* que formam a CDU
- Ex: $\{a_1, b_7, c_8\}$ → CDU
  - O *bin* **a** na dimensão 1 é denso?
  - O *bin* **b** na dimensão 7 é denso?
  - O *bin* **c** na dimensão 8 é denso?
  - Se todas as respostas forem sim → CDU é uma unidade densa

Cada *bin* em cada dimensão tem um limiar diferente

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os bins encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
    Se (K > 1) {Encontre-unidades-densas-candidatas();}
    Leia os dados em blocos de B registros e para cada registro popule as unidades candidatas densas
    Identifique-unidades-densas():
    Registre unidades não densas
fim

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os bins encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
    Se (K > 1) {Encontre-unidades-densas-candidatas();}
    Leia os dados em blocos de B registros e para cada registro popule as unidades candidatas densas
    Identifique-unidades-densas():
    Registre unidades não densas
    Construa-estrutura-dados-unidades-densas();
fim

MAFIA

N – número de registros d – dimensionalidade dos dados
A_i – *i-ésimo* atributo B – número de registros que cabem na memória

Leia os dados em blocos de B registros e
Construa um histograma em cada dimensão Ai, i∈ (1,..,d)
Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i∈ d e determine o nível do limiar
Defina as unidades candidatas densas como os bins encontrados em cada dimensão
Defina a dimensionalidade corrente, k, com o valor 1
Enquanto (forem encontradas mais unidades densas)
    Se (K > 1) {Encontre-unidades-densas-candidatas();}
    Leia os dados em blocos de B registros e para cada registro popule as unidades candidatas densas
    Identifique-unidades-densas():
    Registre unidades não densas
    Construa-estrutura-dados-unidades-densas
Fim
relatorio-grupos();

MAFIA

- Grupos que são um subconjunto de um grupo de maior dimensão são eliminados e somente os grupos de mais alta dimensionalidade são mostrados ao usuário

Paralelismo

- Problema
  - Unir unidades de menor dimensão para formar unidades de alta dimensão requer múltiplos passos sobre os dados
  - Requisitos computacionais aumentam em grandes bases de dados
- Proposta
  - Processamento paralelo → *pMAFIA*

pMAFIA

N – número de registros

p – número de processadores

d – dimensionalidade dos dados

Ai – *i-ésimo* atributo

B – número de registros que cabem no buffer de memória de cada processador

← Termos usados

/* cada processador lê N/p registros para seu disco local */

Em cada processador

Leia N/pB blocos de B registros do disco local e construa um histograma em cada dimensão Ai, i ∈ (1,..., d)

Reduce → comunicação para obter o histograma global

Determine os intervalos adaptativos usando o histograma em cada dimensão Ai, i ∈ d e determine o nível do limiar

Defina as unidades candidatas densas como os *bins* encontrados em cada dimensão

Defina a dimensionalidade corrente, k, com o valor 1

Enquanto (forem encontradas mais unidades densas)

Se (K > 1) {Encontre-unidades-densas-candidatas();}

Leia N/pB blocos de B registros e para cada registro popule as CDU's

Reduce→ comunicação para encontrar unidades densas candidatas globais

Identifique-unidades-densas():

Registre unidades não densas

Construa-estrutura-dados-unidades-densas();

Fim

Se (processador principal)

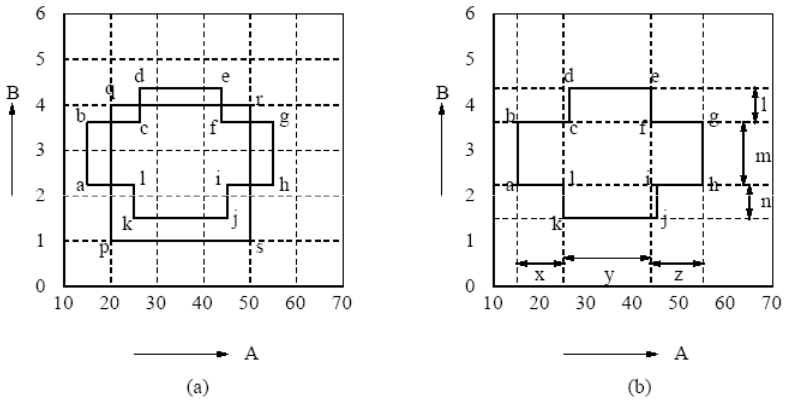
Imprima-grupos();

Fim

Observações

- Os grupos encontrados são representados por expressões DNF

Detalhes do *grid* adaptativo



(a) Cluster discovered by CLIQUE (b) Cluster discovered by MAFIA

Expressão DNF da Figura b: (l,y) ^ (m,z) ^ (n,y) ^ (m,x) ^ (m,y)

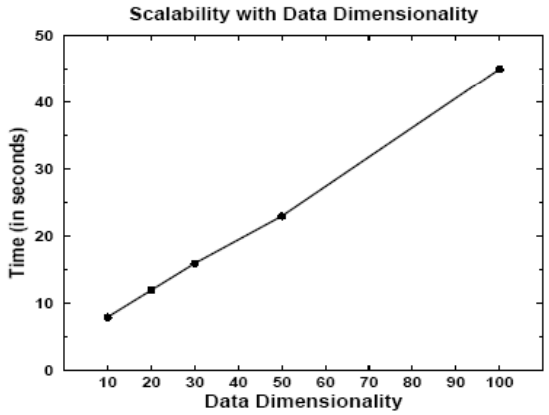
Figura extraída de Nagesh (2001)

Detalhes dos testes - *grid* adaptativo

- Os experimentos usaram
  - Tamanho de cada *bin* inicial em cada dimensão: max(1000, (n-m))
    - n e m é o intervalo da dimensão
  - Janelas de tamanho 5
  - Valor de β: 20%
    - Usado para unir janelas adjacentes
    - Altos valores → tendência a unir todos os *bins*
      - Grupos com qualidade pobre
    - Baixos valores → grande número de *bins*
  - Valor de α: valores maiores que 1,5

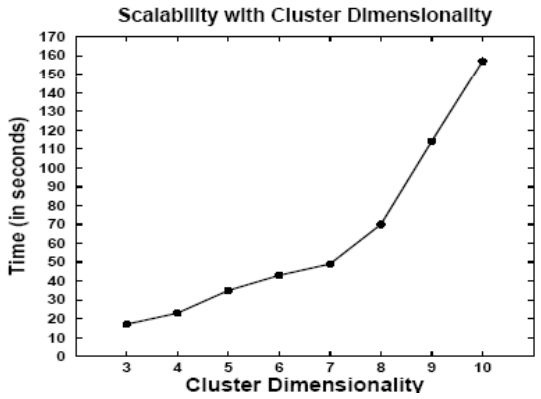
Detalhes dos testes

- Dimensionalidade dos dados



Detalhes dos testes

- Dimensionalidade dos grupos



Detalhes dos testes

- Dados Teste
  - Um único grupo com 7-dimensões
  - Dados: 10 dimensões
  - Registros: 5,4 milhões
- Resultados
  - MAFIA: encontrou o único cluster
  - CLIQUE (modificado): descobriu 75 clusters de 6 dimensões e 546 de 7 dimensões

Referências Extras

- Leung K. e Leckie C., **Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters**, Proceedings of the Twenty-eighth Australasian conference on Computer Science, ACSC '05**,** Volume 38, pp. 333-342, 2005.