CLUTO. A clustering toolkit [1]

Adriana Gabriela Ramírez de la Rosa

INAOE
Coordinación de Ciencias Computacionales
Laboratorio de Tecnologías del Lenguaje

8 de octubre de 2009

Contenido

- Introducción
- Descripción de CLUTO
- Opciones principales de CLUTO
- Información que CLUTO produce
- Ejemplos
- gCLUTO

Introducción

- **Clustering** es la tarea de dividir datos dentro de grupos significativos, útiles o ambos en los cuales poder capturar la estructura de tales datos.
- Propiedades: cohesión interna y separación externa.

Descripción de CLUTO

CLUTO es una familia de programas y librerías de análisis de grupos, adecuados para conjuntos de datos de baja y alta dimensionalidad.

- Características:
  - 3 clases de algoritmos de agrupamiento
  - Opera sobre la matriz de características o sobre una matriz de similitudes
  - Diferentes herramientas de visualización
  - Principal: todos los algoritmos implementados tratan el problema del agrupamiento como un proceso de optimización (maximizar o minimizar una *función de criterio de agrupamiento*)

Opciones principales de CLUTO

1. Control de varios aspectos de los algoritmos de agrupamiento
2. Control del tipo de análisis y reportes generados sobre el agrupamiento
3. Control de la visualización de los grupos

1. Opciones sobre los algoritmos de agrupamiento

- **Tipo de algoritmo**: paradigmas particional, aglomerativo y grafos particionados.
- **Función de similaridad**: coseno, coeficiente de correlación, distancia euclidiana y coeficiente jaccard.
- **Función de criterio de agrupamiento**

Table 1: Clustering Criterion Functions.

$\mathcal{I}_1$	maximize	$\sum_{r=1}^{k} n_r \left(\frac{1}{n_r^2} \sum_{d_i, d_j \in S_r} \cos(d_i, d_j) \right) = \sum_{r=1}^{k} \frac{\|D_r\|^2}{n_r}$	(1)
$\mathcal{I}_2$	maximize	$\sum_{r=1}^{k} \sum_{d_i \in S_r} \cos(d_i, C_r) = \sum_{r=1}^{k} \|D_r\|$	(2)
$\mathcal{E}_1$	minimize	$\sum_{r=1}^{k} n_r \cos(C_r, C) = \sum_{r=1}^{k} n_r \frac{{D_r}^t D}{\|D_r\|}$	(3)
$\mathcal{H}_1$	maximize	$\frac{\mathcal{I}_1}{\mathcal{E}_1} = \frac{\sum_{r=1}^{k} \|D_r\|^2/n_r}{\sum_{r=1}^{k} n_r {D_r}^t D/\|D_r\|}$	(4)
$\mathcal{H}_2$	maximize	$\frac{\mathcal{I}_2}{\mathcal{E}_1} = \frac{\sum_{r=1}^{k} \|D_r\|}{\sum_{r=1}^{k} n_r {D_r}^t D/\|D_r\|}$	(5)
$\mathcal{G}_1$	minimize	$\sum_{r=1}^{k} \frac{\text{cut}(S_r, S - S_r)}{\sum_{d_i, d_j \in S_r} \cos(d_i, d_j)} = \sum_{r=1}^{k} \frac{{D_r}^t (D - D_r)}{\|D_r\|^2}$	(6)

1. Opciones sobre los algoritmos de agrupamiento

Otras opciones interesantes:

- Método para seleccionar el siguiente grupo a dividir:
  - el mas grande,
  - el que maximice (o minimice) la función de criterio de agrupamiento seleccionada, y
  - el que mas reduzca la dimensionalidad del espacio de características.
- *Podado de columnas*. Se indica qué tanto cada columna o característica debe contribuir a la similitud entre las instancias.

2. Opciones para analizar y generar los reportes

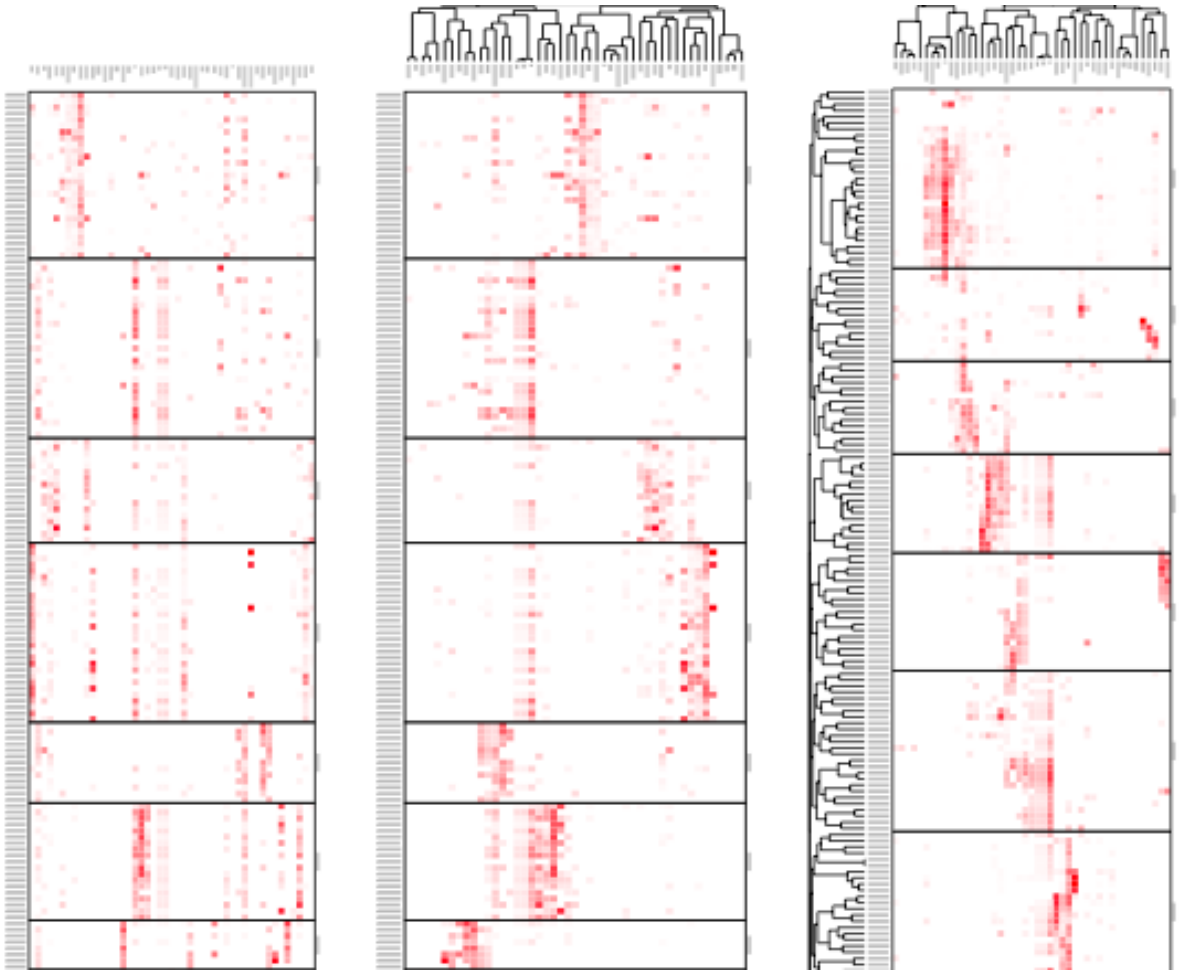
- *showfeatures*. Identifica el conjunto de características que son mas descriptivos de cada grupo y el conjunto de características que mejor discriminan a cada grupo del resto
- *showsummaries*. Identifica relaciones en el conjunto de características descriptivas de cada grupo. La idea es identificar sub-grupos.

2. Opciones para analizar y generar los reportes

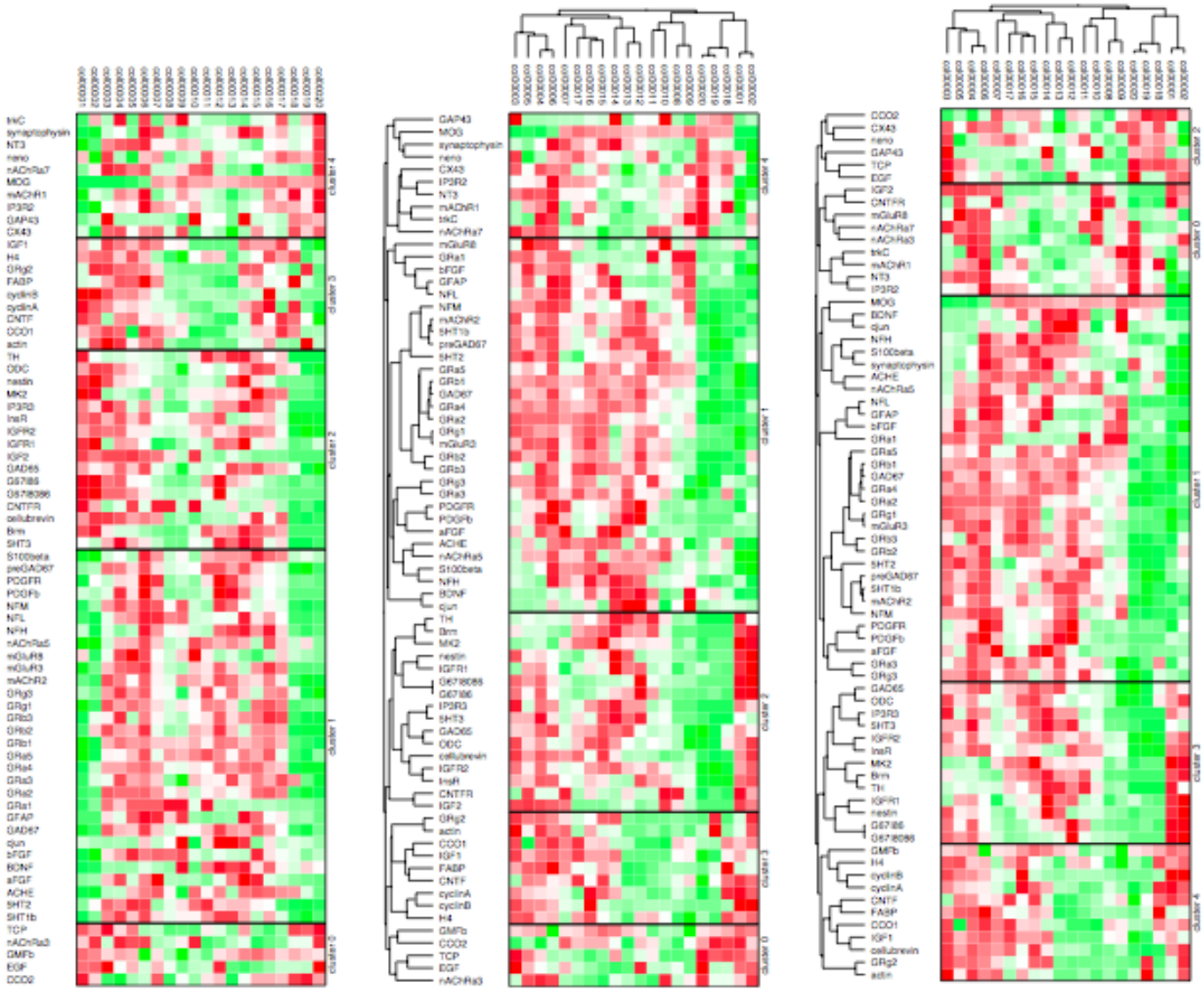
```
10-way clustering solution - Descriptive & Discriminating Features...
================================================================================
Cluster   0, Size:   359, ISim: 0.168, ESim: 0.020
      Descriptive:  warrior 38.1%, hardawai  6.9%, mullin  6.1%, nelson  4.4%, richmond  4.2%
   Discriminating:  warrior 26.6%, hardawai  4.9%, mullin  4.3%, richmond  2.9%, g  2.7%

Cluster   1, Size:   629, ISim: 0.106, ESim: 0.022
      Descriptive:  canseco  9.0%, henderson  7.5%, russa  6.3%, la  3.8%, mcgwire  3.2%
   Discriminating:  canseco  7.5%, henderson  5.9%, russa  5.3%, la  2.6%, mcgwire  2.6%

Cluster   2, Size:   795, ISim: 0.102, ESim: 0.018
      Descriptive:  shark 22.3%, goal  9.4%, nhl  4.4%, period  3.4%, penguin  1.6%
   Discriminating:  shark 17.1%, goal  5.9%, nhl  3.4%, period  2.3%, giant  1.5%

Cluster   3, Size:   762, ISim: 0.099, ESim: 0.021
      Descriptive:  yard 35.8%, pass  7.7%, touchdown  6.5%, td  2.6%, kick  2.1%
   Discriminating:  yard 28.2%, pass  5.4%, touchdown  5.1%, td  2.1%, kick  1.5%

Cluster   4, Size:   482, ISim: 0.098, ESim: 0.022
      Descriptive:  laker  6.0%, nba  3.4%, bull  3.0%, rebound  2.9%, piston  2.5%
   Discriminating:  laker  4.9%, nba  2.7%, bull  2.5%, piston  2.2%, jammer  2.1%

Cluster   5, Size:   844, ISim: 0.095, ESim: 0.023
      Descriptive:  giant 20.7%, mitchell  4.8%, craig  3.3%, mcgee  2.4%, clark  2.0%
   Discriminating:  giant 15.6%, mitchell  4.3%, craig  2.5%, mcgee  2.2%, yard  1.9%

Cluster   6, Size:  1724, ISim: 0.059, ESim: 0.022
      Descriptive:  in  5.6%, hit  5.2%, homer  2.6%, run  2.4%, sox  2.2%
   Discriminating:  in  4.1%, hit  3.4%, yard  2.8%, sox  2.1%, homer  1.8%

Cluster   7, Size:  1175, ISim: 0.051, ESim: 0.021
      Descriptive:  seifert  3.2%, bowl  3.2%, montana  3.1%, raider  2.5%, super  2.0%
   Discriminating:  seifert  3.6%, montana  3.3%, bowl  3.0%, raider  2.5%, super  2.2%

Cluster   8, Size:   853, ISim: 0.043, ESim: 0.019
      Descriptive:  confer  2.4%, school  2.3%, santa  2.1%, st  1.8%, coach  1.8%
   Discriminating:  giant  2.1%, school  1.9%, confer  1.9%, santa  1.7%, yard  1.5%

Cluster   9, Size:   957, ISim: 0.032, ESim: 0.015
      Descriptive:  box 12.4%, golf  3.9%, hole  2.9%, round  2.4%, par  2.0%
   Discriminating:  box  7.6%, golf  3.7%, hole  2.6%, par  1.9%, round  1.5%
--------------------------------------------------------------------------------
```

3. Opciones para visualización

- Utilizando intensidades de colores para indicar los valores de cada característica en una instancia dada.



3. Opciones para visualización



Información producida por CLUTO

Calidad interna. Medida por la función de criterio de agrupamiento usada y la similitud de los objetos en cada grupo

- ISim. Promedio de similitud entre los objetos de cada grupo
- ISdev. La desviación estándar interna
- ESim. Promedio de similitud de los objetos de cada grupo y el resto de los objetos
- ESdev. La desviación estándar de las similitudes externas.

Calidad externa.

- Entropía y Pureza.

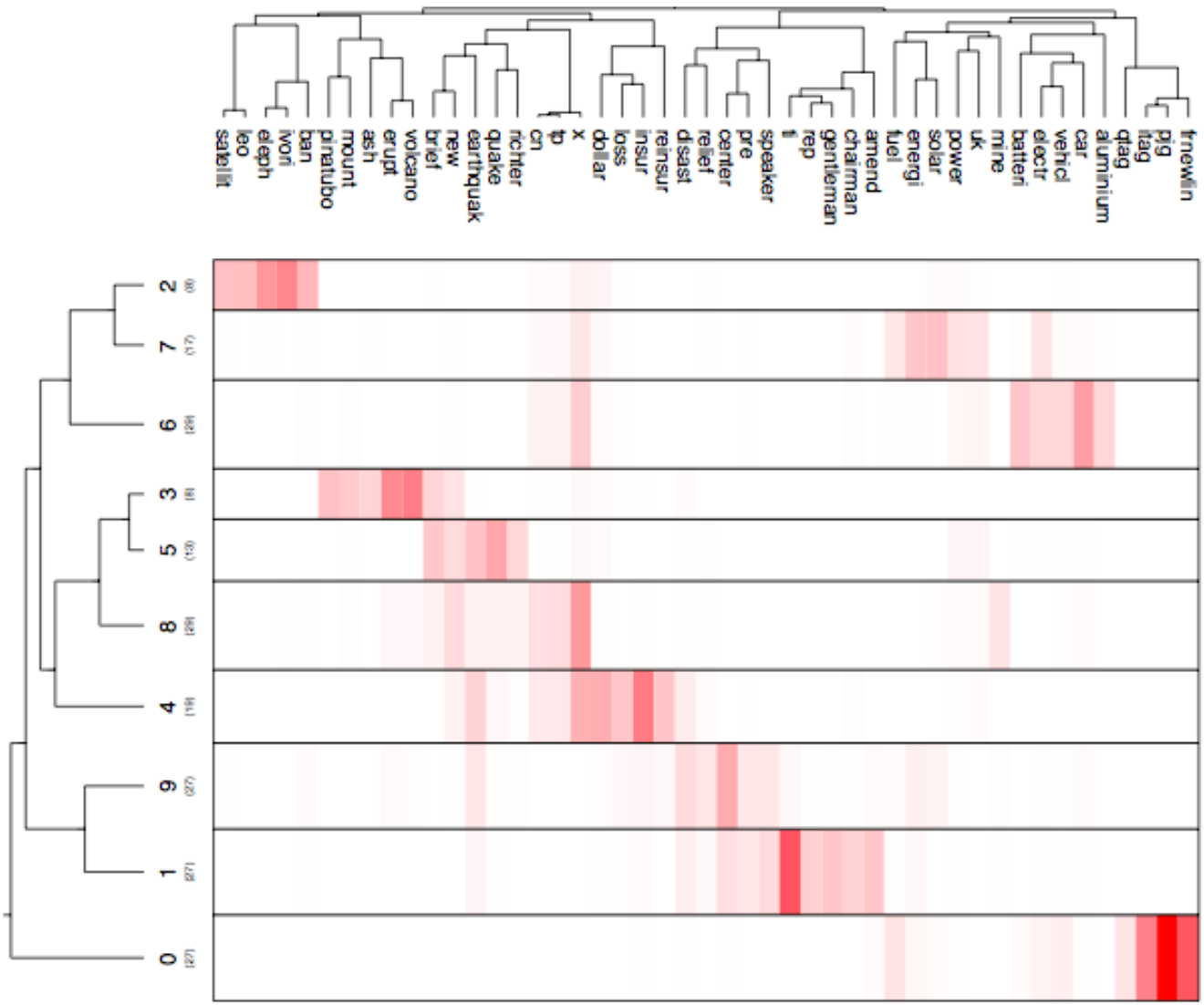
Otros ejemplos

```
prompt% vcluster -rclassfile=sports.rclass sports.mat 10
*******************************************************************************
vcluster (CLUTO 2.1) Copyright 2001-02, Regents of the University of Minnesota

Matrix Information ----------------------------------------------------------
  Name: sports.mat, #Rows: 8580, #Columns: 126373, #NonZeros: 1107980

Options ---------------------------------------------------------------------
  CLMethod=RB, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution --------------------------------------------------------------------

-------------------------------------------------------------------------------
10-way clustering: [I2=2.29e+03] [8580 of 8580], Entropy: 0.155, Purity: 0.885
-------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | base bask foot hock boxi bicy golf
-------------------------------------------------------------------------------
  0   359 +0.168 +0.050 +0.020 +0.005 0.010 0.997 |    0  358    1    0    0    0    0
  1   629 +0.106 +0.041 +0.022 +0.007 0.006 0.998 |  628    0    1    0    0    0    0
  2   795 +0.102 +0.036 +0.018 +0.006 0.020 0.995 |    1    1    1  791    0    0    1
  3   762 +0.099 +0.034 +0.021 +0.006 0.010 0.997 |    0    1  760    0    0    0    1
  4   482 +0.098 +0.045 +0.022 +0.009 0.015 0.996 |    0  480    1    1    0    0    0
  5   844 +0.095 +0.035 +0.023 +0.007 0.023 0.993 |  838    0    5    0    1    0    0
  6  1724 +0.059 +0.026 +0.022 +0.007 0.016 0.996 | 1717    3    3    1    0    0    0
  7  1175 +0.051 +0.015 +0.021 +0.006 0.024 0.992 |    8    1 1166    0    0    0    0
  8   853 +0.043 +0.015 +0.019 +0.006 0.461 0.619 |   46  528  265    8    0    0    6
  9   957 +0.032 +0.012 +0.015 +0.006 0.862 0.343 |  174   38  143    8  121  145  328
-------------------------------------------------------------------------------

Timing Information ----------------------------------------------------------
   I/O:                                    1.620 sec
   Clustering:                             9.110 sec
   Reporting:                              0.230 sec
*******************************************************************************
```

Otros ejemplos

```
--------------------------------------------------------------------------------
Hierarchical Tree that optimizes the I2 criterion function...
--------------------------------------------------------------------------------
                          base bask foot hock boxi bicy golf
----------------------------------------------------------------
18
|-----15
|      |-------------6     1717    3    3    1    0    0    0
|      |---13
|          |---------1      628    0    1    0    0    0    0
|          |---------5      838    0    5    0    1    0    0
|-17
  |---------12
  |          |-------7        8    1 1166    0    0    0    0
  |          |-------3        0    1  760    0    0    0    1
  |-16
    |---14
    |    |-----11
    |    |      |-----8       46  528  265    8    0    0    6
    |    |      |-----9      174   38  143    8  121  145  328
    |    |-------10
    |             |---0        0  358    1    0    0    0    0
    |             |---4        0  480    1    1    0    0    0
    |---------------2          1    1    1  791    0    0    1
----------------------------------------------------------------
--------------------------------------------------------------------------------
```

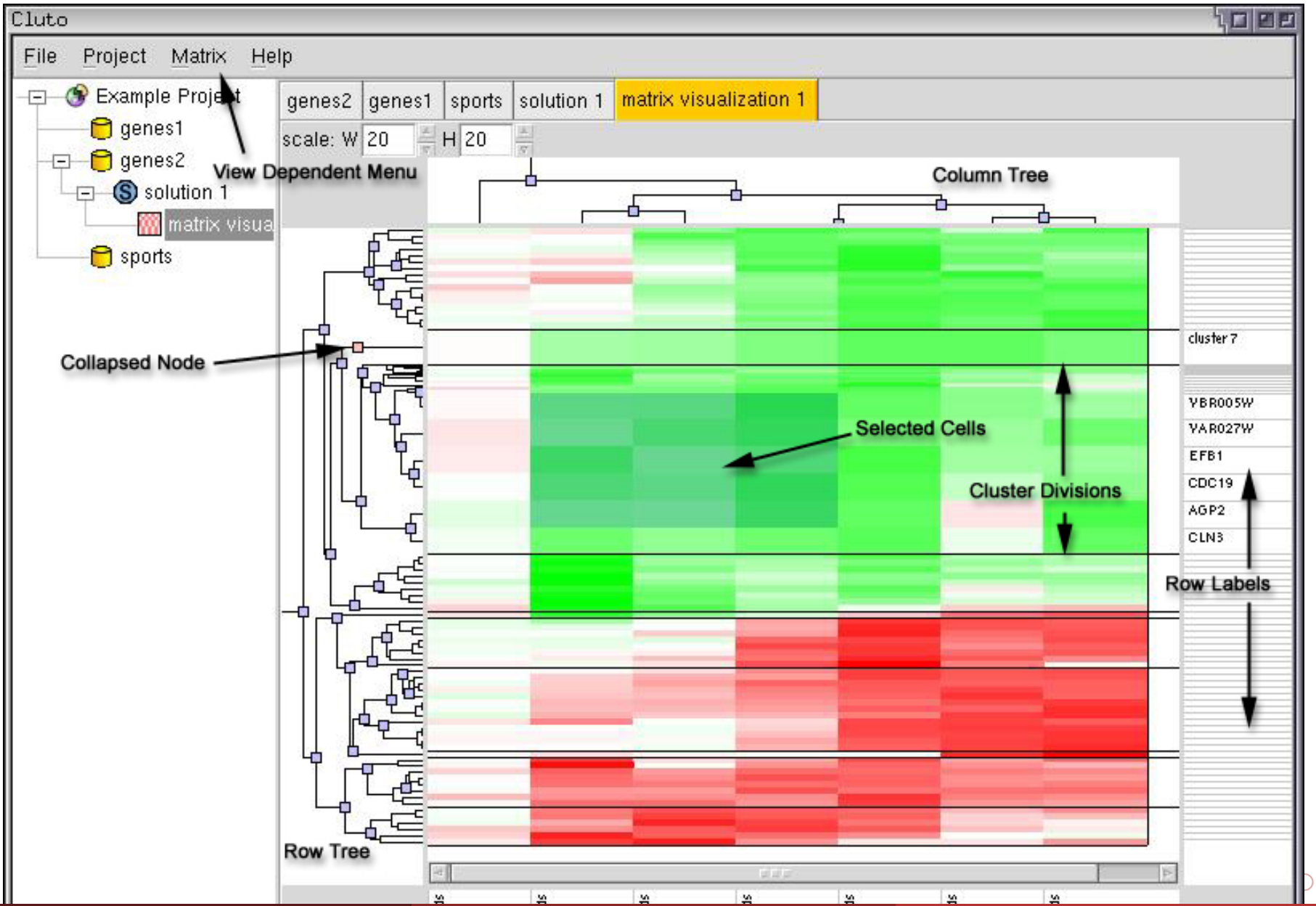
Otros ejemplos



gCLUTO


Cluto
File Project Mountain Window Help
myproject
genes1
solution 1
solution 2
matrix visuali
matrix visuali
solution 3
k1b
solution 1
matrix visuali
sports
a
solution 1
solution 1 - Solution Resu
Clustering Options
Method: Repeated Bisection
#Clusters: 10
CRfun: I2
Simfun: Cosine
b
RowModel: None
Col Model: None
Graph Model:
Col Prune: 1.000000
EdgePrune: 0.500000
VertexPrune:
Nearest Nieghbors: 4
MinComponent: 1
CSType: Best
#Trials: 10
#Iterations: 10
10-way clustering solution
cluster size ISim ISdev ESim
0 1 1.000000 0.000000 0.747020
1 2 0.984784 0.000000 0.751680
matrix visualization 1
scale: W 27 H 22
c
mountain visualization 1
d
0
1
2
3
4
5
6
7
8
9

gCLUTO


Cluto
File
Project
Matrix
Help
Example Project
genes1
genes2
solution 1
matrix visua
sports
genes2
genes1
sports
solution 1
matrix visualization 1
scale: W 20
H 20
View Dependent Menu
Column Tree
Collapsed Node
Selected Cells
Cluster Divisions
Row Labels
Row Tree
cluster 7
YBR005W
YAR027W
EFB1
CDC19
AGP2
CLN3

Referencias I

Karypis, G.
CLUTO-A Clustering Toolkit, 2002.

¡GRACIAS!

¿Alguna pregunta, opinión o sugerencia?