·技术与方法·

MATLAB 7.X 生物信息工具箱的应用——聚类分析(七)*

刘新星 梁万洁 李红燕 杨英杰△

(中南大学生物冶金教育部重点实验室 湖南 长沙 410083)

摘要 本文主要介绍 MATLAB 生物信息工具箱的数据聚类分析功能,该功能主要用于基因芯片数据的分析。将要分析的数据先转化成 XLS 格式的文件,通过函数 xlsread 读入 MATLAB Workspace,存储为两个变量。对缺失数据进行估算,从而减小结果误差。函数 clustergram 对数据分级聚类,并产生数据的热红外分布图和树状图。通过更改相关参数可以改变其颜色配置,距离算法,并可做双向聚类。

关键词 聚类分析;电子数据表;Clustergram;基因芯片

中图分类号:TP391 **文献标识码**:B **文章编号**:1673-6273(2012)17-3259-04

The Application of MATLAB Bioinformatics Toolbox--Cluster Analysis(7)*

LIU Xin-xing, LIANG Wan-jie, LI Hong-yan, YANG Ying-jie△

(Key Laboratory of Biometallurgy of Ministry of Education, Central South University, Changsha, 410083, China)

ABSTRACT: This paper introduces the function of MATLAB Bioinformatics Toolbox supplies functions on data clustering analysis, which is used mostly to analyze the gene chip data. The data should be converted into XLS files at first, then read into the MATLAB Workspace by the function of xlsread and stored as two variables. Inputing values for defaults is helpful for diminishing resultant errors. The function clustergram is used to perform hierarchical clustering and to generate a heat map and a dendrogram of the data. By changing the relevant parametres , the color scheme and the distance arithmetic can be changed, and we can also perform biclustering.

Key words: Cluster analysis; Excel spreadsheet; Clustergram; Gene-chip

Chinese Library Classification(CLC): TP391 Document code: B

Article ID: 1673-6273(2012)17-3259-04

前言

聚类分析是一种探索性的数据分析方法。根据目标研究对象的数值属性特征,采用数学方法对之进行分类整理,再对同类个体的共性及差异作进一步的归纳,从而得到新规律[1]。

近年来聚类分析研究发展迅速,从数学、统计学、信息学、人工智能等角度不断有新的方法提出、改进,并在经济学、地质学、气象学、生物学等领域得到成功的应用。在目前生物信息学领域的研究中,聚类分析受到广泛重视[2]。在基因的表达、DNA 序列的研究中,聚类分析已经成为标准的程序。展望生物技术发展的特点,一是将产生数量极为巨大的数据;二是基于这些大量的数据,科研活动将逐步从传统的以实验为主的方式向数据分析与实验相结合的方式过渡。在这一过程中,统计聚类分析将是开展数据分析工作的基石。

MATLAB 生物信息工具箱中包含的 clustergram 函数即用于数据的聚类分析。生物信息工具箱中这一函数主要用于基因表达数据的分析,不仅可横向聚类,还可以纵向聚类。

1 数据预处理

我们使用的数据来自 Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA,Pommier Y, Botstein D, Brown PO, Weinstein JN 的文章《A gene expression database for the molecular pharmacology of cancer》Nature Genetics 2000 Mar;24(3):236-44. PMID: 10700175[3]。

1.1 将数据载入 MATLAB 工作区

例:118 种预测作用的药物应用于 NCI60 细胞系时产生的的生长抑制因素,包含了 118× 60 个数据。原始数据可以从下列网址获得:http://discover.nci.nih.gov/nature2000/data/selected_data/a_matrix118.txt[4]。在本例中,此数据已经被转化为 Excel 的

* 基金资助:国家自然科学基金项目(50774102)

作者简介:刘新星(1955-),女,教授,主要研究方向:生物信息学,E-mail:xinxingliu@hotmail.com

△通讯作者:杨英杰,E-mail:yjyangcsu@126.com

(收稿日期:2012-01-10 接受日期:2012-03-21)

电子数据表。

用函数 xlsread 从电子数据表中读取数据。

```
[numericData, textData] = xlsread('cancerdata.xls');
```

数据即作为 numericData 和 textData 两个变量载入 MATLAB，出现在 workspace 中。

1.2 从 Excel 表提取数据

函数 xlsread 将电子数据表中的数据读取后作为两个变量储存。其中，变量 numericData 储存数值 ;变量 textDate 储存表中的所有文本信息。本例中表格的前三行都是关于实验中药物的文本信息，储存在 textData 变量中。

```
%提取数据，为第二列至最后一列
giValues = numericData(:,2:end);
%提取药物作用机制名称，为文本变量的第一列的第二行至最后一行
drugMechanism = textData(2:end,1);
%提取药物名称，为文本变量的第二列的第二行至最后一行
drugName = textData(2:end,2);
%变量 drug 为变量 drugMechanism 和 drugName 用 - 符号水平连接
drug = strcat(drugMechanism,'-',drugName);
%提取药物的 ID 号，为变量 numericData 的第一列
drugID = numericData(:,1);
%细胞系名称，为第一行的第四列至最后一列
cellLine = textData(1,4:end);
%定义肿瘤细胞类型即细胞系元素以:隔开
tumorTypes = strtok(cellLine,':');
%清除不再需要的变量 numericData、textData
clear numericData textData
```

1.3 估算缺失数据的值

上述程序中的变量 giValues 中包含了一些标志为 NaN 的缺失数据。我们可以简单移除这些数据，也可以估算这些缺失数据[5]。函数 nanmedian 是用来计算行的中间值或列中被忽略的缺失数据。统计学工具箱中用来处理 NaN 的函数还包括 nanmean，nanvar 和 nansum。

```
%用函数 isnan 找出缺失数据
missingVals = isnan(giValues);
%忽略 NaN 值计算行和列的中间值
colMedians = nanmedian(giValues);
rowMedians = nanmedian(giValues,2);
%用中间值替换缺失数据
rowMed = repmat(rowMedians,1,size(giValues,2));
giValues(missingVals) = rowMed(missingVals);
```

2 聚类分析

2.1 聚类分析

函数 clustergram 用于分级聚类，并产生数据的热红外分布图和树状图。聚类的最简单方式是通过相关性的距离度量和平均联系将数据集聚类。本例中，数据进行聚类分析后，其热红外分布图和树状图显示，作用方式相似的药物聚类在一起。

```
clustergram(giValues,'rowlabels',drug,'columnlabels', tumorTypes);
```

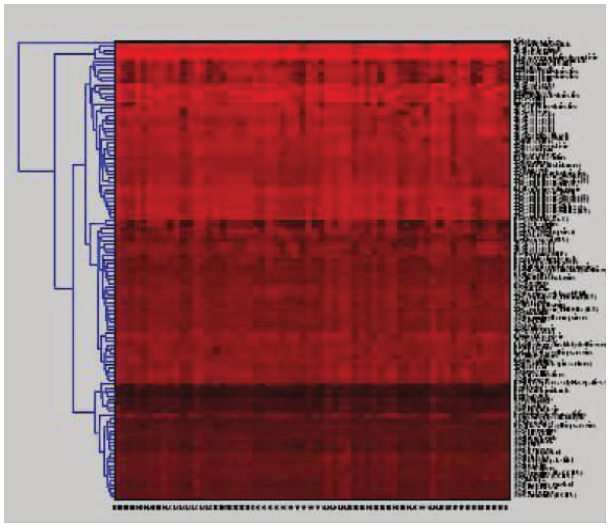

图 1 聚类分析图

Fig.1 Clustering map

2.2 更改颜色配置

默认的比色刻度尺是红 - 绿色，这也广泛应用于微阵列分析中。在本例中，设置不同的比色刻度尺将会更有利于结果的观察。参数 colormap 即可用来指定可供选择的颜色配置。默认的比色刻度尺是基于数据按照 0 对称的，而本例中的数据并不是对称的，因此，设置参数 symmetricRange 为 false，可以在热红外分布图中看到更多的密集范围[6]。

```
clustergram(giValues,'rowlabels',drug,'columnlabels', tumorTypes,'colormap',jet,'symmetricrange',false);
```

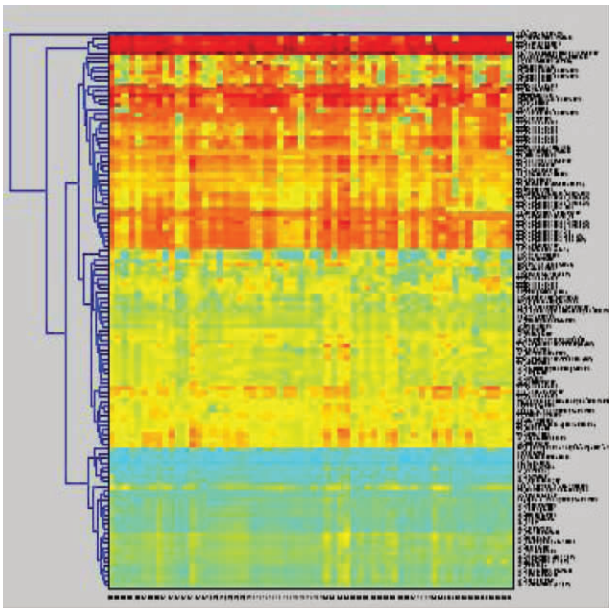

图 2 Jet 颜色方案的聚类分析图

Fig.2 Clustering map using colormap Jet

2.3 数据组横向聚类

进行数据组横向聚类最简单的方法就是利用算子做矩阵转置。虽然，行标签和列标签都转换了，但是树状图却仍然是横向的。

clustergram (giValues,'columnlabels',drug,'rowlabels', tumorTypes,'colormap',jet,'symmetricrange',false);

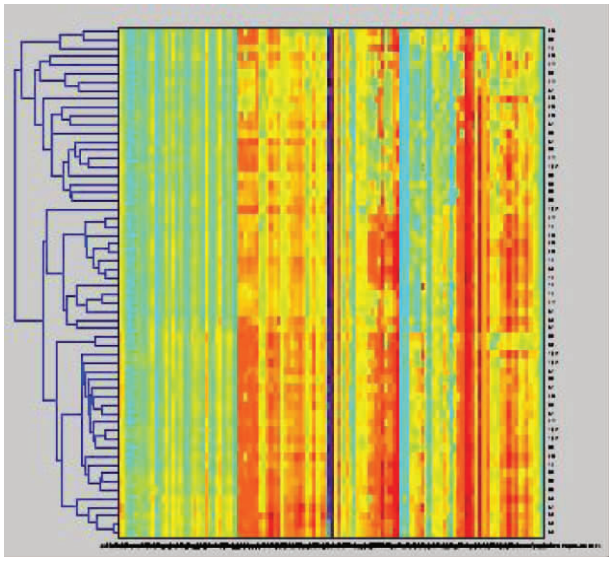

图 3 Jet 颜色方案的横向聚类分析图

Fig.3 Horizontal clustering map using colormap Jet

2.4 双向聚类分析

做双向聚类分析，需要将参数 dimension 设定为 2。这种聚类分析的结果同时包含了数据的行和列，并绘制包含了两个方向的树状图和热红外分布图，其中一个树状图显示细胞系的聚类结果，另一个树状图则显示药物的聚类结果。

clustergram(giValues,'dimension',2,'rowlabels',drug,'columnlabels', tumorTypes,'colormap', jet,'symmetricrange', false);

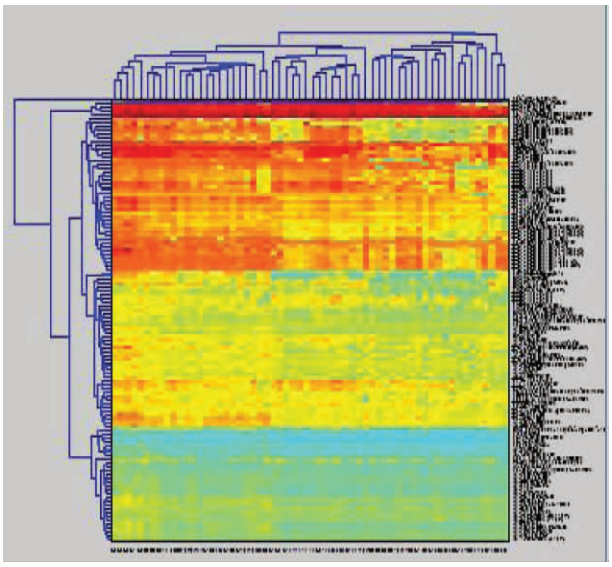

图 4 Jet 颜色方案的双向聚类分析图

Fig.4 Bi-clustering map using colormap Jet

2.5 更改聚类分析的参数

可以通过更改聚类分析算法的参数以使用不同的距离计算方法或联系方式进行聚类。本例中，我们改用加权联系（WPGMA）计算距离[7]，并设定树状图可以突出距离小于 10 单位的聚类群。

clustergram(giValues,'dimension',2,'rowlabels',drug,'columnlabels',tumorTypes,'colormap',jet,'symmetricrange',false,'linkage','weighted','dendrogram',{'color',10});

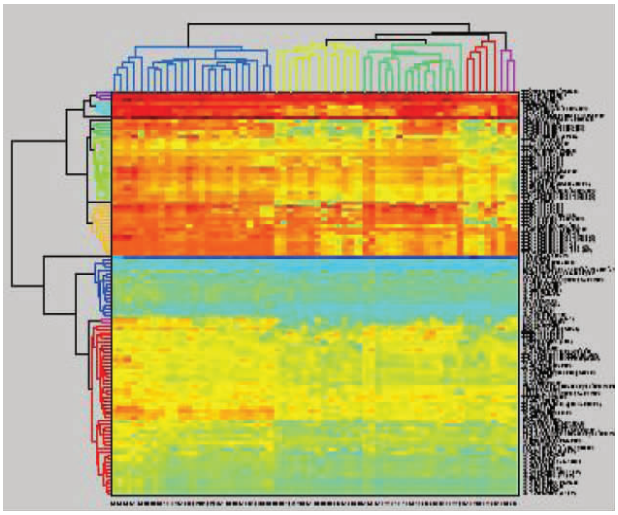

图 5 加权联系的双向聚类分析图

Fig.5 Bi-clustering map by WPGMA

3 小结

MATLAB 生物信息工具箱中的函数很丰富，用法也很多，功能比较强大，对大量所得数据进行分析时是一个很好的帮手。以上介绍的都只是 MATLAB 生物信息工具箱中最基本的函数功能和用法，更多的内容则需要通过进一步查询和使用才能熟悉。在完全掌握了众多函数的使用方法后，MATLAB 能发挥最佳的数据分析功能。

参考文献(References)

[1] Richmond C.S., Glasner J.D., Mau R., et al. Genome-wide expression profiling in Escherichia coli K-12. Nucleic Acids Research, 1999, 27: 3821-3835

[2] 刘月明. 基因表达数据聚类分析方法研究[D]. 第三军医大学, 2001 Liu Yue-ming. Clustering Analysis of Gene EzpressionData [D]. The Third Military Medical University, 2001

[3] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA,Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. Nature Genetics, 2000, 24 (3):236-44. PMID: 10700175

[4] http : //discover . nci . nih . gov/nature2000/data/selected_data/a_matrix 118.txt

[5] 常世杰,尹勇,龙哲,沙宪政. Matlab 在生物信息分析中的应用前景[J]. 生物医学工程研究, 2006, 3: 186-190

Chang Shi-jie, Yin Yong, Long Zhe, et al. A Vision for Matlab in Bioinformatics Research [J]. Journal of Biomedical Engineering Research. 2006, 3: 186-190

[6] The MathWorks. Bioinformatics toolbox for use with MATLAB[M]. The MathWorks Inc, 2005: 1-2

[7] 包志强, 吴顺君, 韩冰. 一种广义加权模糊聚类算法[J]. 华中科技大学学报(自然科学版), 2007 年 S1 期

Bao Zhi-qiang, Wu Shun-jun, Han Bing. A general weighted fuzzy clustering algorithm [J]. Journal of Huazhong University of Science and Technology(Nature Science Edition, 2007(S1)

(上接第 3231 页)

[19] 胡辉,宋义军. 西红花的药学、药理学及其应用概述[J].新疆中医药, 2005,23(4):72-74

Hu Hui, Song Yi-jun. The Outline of Xi Hong Hua that pharmacy pharmacology and its application[J]. Traditional Chinese Medicine of Xinjiang, 2005, 23(4):72-74

[20] 汪云,李红霞. 藏红花对大鼠肝毒性的实验研究[J]. 哈尔滨医科大学校报, 2010, 02:133-135

Wang Yun, Li Hong-xia. The experimental study of liver toxicity in rats with zang Hong Hua[J]. The school newspaper of Harbin Medical University, 2010,02:133-135

[21] 杨春潇,李丽丽.藏红花对 CCL4 致小鼠急性肝损伤的保护作用[J]. 现代中医药杂志, 2009, 02:64-65

Yang Chun-xiao, Li Li-li. The protection of Acute liver injury in rats by CCL4[J]. Modern Chinese Medicine magazine, 2009,02:64-65

[22] Tsukada S, Westwick JK, IkejimaK, et al. SMAD and p38MAP signaling pathways independently regulate(I) collagen gene expression in unstimulated and trans-forming growth factor stimulated hepatic stellate cells[J]. J BiolChem, 2005, 280(11):10055-10064