

一种 CCA - 层次聚类的基因聚类算法

林倩闽

(厦门理工学院 电气工程与自动化学院, 福建 厦门 361024)

摘要:针对基因芯片技术带来的海量基因表达数据,为了充分挖掘其蕴含的生物信息和潜在的生物机制,提出一种基于 CCA - 层次聚类的基因聚类算法(CCA-Hc)。该算法在层次聚类的基础上引入典型相关分析,优化相似性矩阵计算方法。首先,利用典型相关分析方法结合基因的多个特征信息进行基因相关性度量,得到基因相似性矩阵。然后将该相似性矩阵作为层次聚类的邻近矩阵进行凝聚层次聚类。在 *Oryza sativa* L. (水稻)的基因表达数据集上进行 CCA-Hc 聚类效果测试实验,结果表明,与采用欧式距离的传统层次聚类算法(EUC-Hc)相比,CCA-Hc 的内部稳定性指标和生物功能性指标均优于 EUC-Hc,具有更佳的鲁棒性和聚类准确性,更有利于去发现基因间的共表达关系。

关键词:基因表达数据;聚类算法;典型相关分析;层次聚类

DOI:10.15938/j.jhust.2023.05.011

中图分类号: TP391 **文献标志码:** A **文章编号:** 1007-2683(2023)05-0085-06

A Gene Clustering Algorithm Based on the CCA-Hierarchical Clustering

LIN Qianmin

(School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China)

Abstract: Aiming at the massive gene expression data brought by gene chip technology, in order to fully mine the biological information and potential biological mechanisms contained in it, this paper proposes a gene clustering algorithm based on CCA-hierarchical clustering (CCA-Hc). The algorithm introduces canonical correlation analysis on the basis of hierarchical clustering, and optimizes the calculation method of similarity matrix. First, the canonical correlation analysis method is used to measure the gene correlation by combining the multiple feature information of the gene, and the gene similarity matrix is obtained. Then the similarity matrix is used as the neighbor matrix of hierarchical clustering for agglomerative hierarchical clustering. The CCA-Hc clustering effect test experiment was performed on the gene expression dataset of *Oryza sativa* L. (rice). The results show that, compared with the traditional hierarchical clustering algorithm using Euclidean distance (EUC-Hc), CCA-Hc is superior to EUC-Hc in both internal stability index and biological functional index, and has better robustness and clustering accuracy. It is more conducive to discovering the co-expression relationship between genes.

Keywords: gene expression data; clustering algorithm; canonical correlation analysis; hierarchical clustering

0 引言

随着高通量测序技术的不断快速发展,出现越

来越多复杂度高、数据量大的生物数据。不同测序技术可以得到不同水平的生物数据,如通过基因组测序得到 DNA 水平的生物数据,转录组测序得到 RNA 水平的生物数据。基因表达数据是通过 DNA

微阵列技术(又称为基因芯片技术)检测得到,是不同细胞在不同条件下的基因动态表达水平^[1]。基因是携带遗传物质的 DNA 片段,在不同细胞中会有不同的表达方向^[2],从而可以控制不同的性状。为此基因表达数据蕴含着丰富且重要的生物机制,具有很大的研究价值。

在基因表达数据分析中,聚类分析方法被广大研究者选用,用以发现具有相似表达行为的基因集,基因间的共表达、共调控关系等,对于推断未知的基因功能及在疾病诊断方面具有重要意义^[2]。目前基因聚类算法根据聚类对象可以分为基于基因、基于样本聚类以及基于基因样本的双聚类^[3-4]。根据聚类方式的不同,又可以分为以 K-means 算法^[5]、K-MEDOIDS^[6]为代表的基于分区的聚类算法,以 BIRCH 算法^[7]、CURE 算法^[8]为代表的基于层次的聚类算法,以 DBSCAN 算法^[9]、OPTICS 算法^[10]为代表的基于密度的聚类算法和以 CLIQUE 算法^[11]为代表的基于网格的聚类算法。

在对基因表达数据进行聚类分析时,主要是度量基因之间的相关性,把相关性程度高的基因聚在一起。很多基因聚类研究中把皮尔森相关系数、欧式距离、曼哈顿距离等作为相关性程度的度量方式^[12]。这些度量方式是基于基因的整体表达水平进行的,即一个基因只由一个一维的数据矩阵表示。而在实际的测序过程中,往往会在不同的细胞周期进行实验测量基因的表达水平,使得一个基因会有多组数据,每组数据代表该基因的一个特征。大部分的研究中采用求和的方式把基因多个特征的数据进行累加,进而分析基因之间的相关性。这种方法存在的问题是忽略了基因各个特征对表达水平的影响,从而对聚类结果造成影响。

为了解决上述问题,本文把典型相关分析(Canonical Correlation Analysis, CCA)引入到层次聚类中来,搭建出基于 CCA-层次聚类的基因聚类算法(CCA-Hc)。典型相关分析是一种计算变量之间相关性的统计学分析方法,能结合变量的多个特征,得到变量的整体相关性^[13]。利用典型相关分析度量基因之间的相关性,能充分考虑基因的多个特征信息,使得聚类结果中的基因集相似性程度更高。同时采用凝聚层次聚类,可以从聚类树状图中直观地分析聚类结果,从而整体上提高聚类效果。最后用

GEO 数据库上的基因数据集来验证 CCA-Hc 算法的有效性。

1 CCA-Hc 算法设计

1.1 典型相关分析

给定基因微阵列数据矩阵 $A_{n \times m} = (G, T)$, n 表示基因个数, m 表示条件的种类数。每个基因可以看成是一个变量,使用典型相关分析方法分析变量相关性时,假设变量 X 有 p 个特征,变量 Y 有 q 个特征, $p \leq q$, 每个特征均对应 m 个不同条件的数据,则

$$X = [x_1, \dots, x_p]^T \quad (1)$$

$$Y = [y_1, \dots, y_q]^T \quad (2)$$

变量 X 的数据矩阵为

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & x_{p3} & \cdots & x_{pm} \end{bmatrix}$$

变量 Y 的数据矩阵为

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1m} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2m} \\ y_{31} & y_{32} & y_{33} & \cdots & y_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{q1} & y_{q2} & y_{q3} & \cdots & y_{qm} \end{bmatrix}$$

变量 X 和变量 Y 的协方差矩阵为

$$\Sigma = \text{Cov}(X, Y) = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (3)$$

变量 X 和变量 Y 的线性表达式记为 U, V , 表示为:

$$U = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p = a^T X \quad (4)$$

$$V = b_1 y_1 + b_2 y_2 + \cdots + b_q y_q = b^T Y \quad (5)$$

变量 X 和变量 Y 进行典型相关性分析时,可用这两个变量的线性表达式 U, V 之间相关系数的最大值来度量变量之间的相关性程度,即

$$\max_{a, b} \text{corr}(U, V) = \frac{a^T \Sigma_{12} b}{(a^T \Sigma_{11} a \times b^T \Sigma_{22} b)^{1/2}} \quad (6)$$

在求解上述最值表达式时,运用拉格朗日数乘法求解瑞利熵矩阵 $(\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$ 得到 p 个特征值,

记为 $\lambda_1, \lambda_2 \cdots \lambda_p$ 。这 p 个特征值即变量 X 和变量 Y 之间的典型相关系数。每一个相关系数再应用卡方检验进行显著性检验,得到 p 个卡方检验 p -value 值,记为 $p_1, p_2 \cdots p_p$ 。为了更好地表示变量之间的典型相关程度,引入一个关于典型相关系数和 p -value 值的权重函数 W 来表示,定义为:

$$W = \frac{\sum_{i=1}^p \lambda_i I(\log P_i)}{\sum_{i=1}^p I(\log P_i)} \tag{7}$$

其中 $I(\log P_i) = \begin{cases} 0 & P > 0.05 \\ -\log P & P \leq 0.05 \end{cases}$

这样每两个变量之间就能得到一个 w 值来度量它们的相关性程度。对基因表达数据的 n 个基因进行如上方法的典型相关分析后,最终得到一个 $n \times n$ 的相似性矩阵。

1.2 层次聚类

目前常用的聚类算法有基于分区、基于层次、基于密度和基于网络 4 种类型^[2],其中基于层次聚类的算法因原理通俗易懂、结果直观且精度高等优点而被广泛使用^[14]。层次聚类分为自下而上的凝聚聚类 and 自上而下的分裂聚类两种^[15],其中凝聚层次聚类运用最为广泛,同时凝聚层次聚类在无预先定义类别数的分类中具有明显优势^[16]。故本文采用的是凝聚层次聚类,可以用树状图和嵌套簇图来表示,例如图 1 所示。

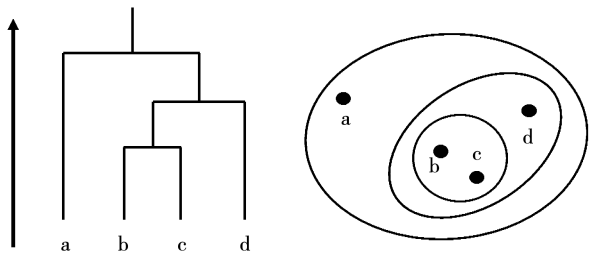


图 1 凝聚层次聚类的树状图和嵌套簇图

Fig.1 Dendrogram and Nested Cluster Diagram for Agglomerative Hierarchical Clustering

下面介绍凝聚层次聚类的聚类过程:

步骤 1:视每一个数据点(如基因变量)为一个集群;

步骤 2:计算邻近矩阵,把类间距离最接近的两个集群进行合并;

步骤 3:重复步骤 2,直到所有数据点合并完成。

步骤 2 中的类间距离即两个集群之间的距离,传统的层次聚类类间距离计算方法有如下几种^[17]:

- 1)两个集群中距离最近的两个样本距离;
- 2)两个集群中距离最远的两个样本距离;
- 3)两个集群中所有样本之间的距离再求平均值;

完成所有聚类步骤后会生产一个树状图(又叫聚类树)。采用不同的变量相关性程度度量方式和不同的类间距离计算方法都将对聚类结果造成影响。

1.3 CCA-HC 算法

传统的层次聚类算法其计算复杂度为 $O(n^3)$,由于在聚类过程中需要不断地重复计算类间距离、不断地更新邻近矩阵,从而消耗大量的时间与资源^[18]。对于数据量庞大的基因微阵列数据,迫切需要对算法进行优化,降低复杂度。本文提出了一种基于 CCA 和层次聚类的基因聚类算法(CCA-HC),优化相似性矩阵计算方法,把典型相关分析的输出作为层次聚类的输入,即把典型相关分析得到的相似性矩阵作为层次聚类的邻近矩阵。

CCA-HC 在度量基因相关性程度时采用典型相关分析的方法,在层次聚类方式上选择自下而上的凝聚层次聚类。CCA-HC 充分利用了典型相关分析和层次聚类的优点,能够结合基因的多个特征来量化基因之间的相关性,使得聚类结果中的基因集相似性程度更高,也能自主选择集群数目以得到最佳的聚类效果^[18]。

2 实验与结果分析

2.1 实验数据

为了评价章节一中提出算法的聚类效果,在 GEO 数据库上下载 *Oryza sativa* L. (水稻)的基因表达数据集,得到的原始数据集共有 45 063 个基因,样本数为 41。由于原始数据集基因数庞大,对其计算分析时不论在存储空间还是计算程序上都提出了较高的要求,为此进行适当的数据预处理显得尤为重要。

本文在数据预处理方面开展的主要工作有:把基因名未知的数据剔除;过滤掉样本表达量过低的基因;采用 $\log_2$ 的对数函数对原始数据进行标准化处理等。经过如上处理后得到 $4\,564 \times 41$ 的数据矩阵,用于后续的实验分析。预处理后的实验数据集

统计情况如表 1 所示。

表 1 预处理后的实验数据集统计情况表
Tab.1 Statistical table of experimental data
set after preprocessing

数据集	基因数	样本数	基因功能类别
Oryza sativa L.	4 564	41	88

1.5 评价标准

基因表达数据的聚类效果可以从聚类结果中同一集群的相关性程度以及聚类算法的稳定性等方面进行评价,用生物功能性指标和内部稳定性指标来描述。

1. 生物功能性指标

生物同源性指标 (biological homogeneity index, BHI) 是用来评估聚类集群在生物功能意义上的同源性程度^[19]。在基因本体 (gene ontology, GO) 数据库上下载水稻的基因功能类数据,可以得知每个水稻基因所对应的生物组织功能,用来分析同一聚类集群中的基因在功能上的相关性。BHI 公式计算如下:

$$BHI(K,B)=\frac{1}{K}\sum_{k=1}^K\frac{1}{n_k(n_k-1)}\sum_{i\neq j\in C_k}I(B(i)=B(j))$$

(8)

式中: C 为聚类结果中的任一集群; B 为基因功能类集合,当基因 i 和基因 j 所对应的功能类存在交集,则 $I(B(i)=B(j))=1$, 否则为 0。最终得到的 BHI 是介于 0~1 的值,BHI 值越大,表示基因聚类集群的生物功能相关性越大,聚类效果更佳^[19]。

2. 内部稳定性指标

内部稳定性指标在于评价聚类算法的鲁棒性,通过改变基因微阵列数据的某几列进行聚类,进而比较基于不同数据的聚类结果。优值系数 (figure of merit, FOM) 是内部稳定性指标中的一种,表示数据列改变后基因之间的平均群内方差^[20]。FOM 公式计算如下:

$$FOM(l,K)=\sqrt{\frac{1}{N}\sum_{k=1}^K\sum_{i\in C_k(l)}dist(x_{i,l},\bar{x}_{C_k(l)})}$$

(9)

式中:FOM 的取值范围是 0 到无穷大,FOM 值越小表示该聚类算法的稳定性越好^[20]。

2.3 结果与分析

为验证 CCA-Hc 的聚类效果,对比采用欧式距

离的传统层次聚类算法 (EUC-Hc),运用相同数据集进行实验。为了获得更加准确的聚类效果,本实验设置不同的聚类集群参数,确定聚类集群数目 K 分别为 2、4、6、7、9、11、12 这 7 组实验,并通过 BHI 和 FOM 指标对这 7 组实验的聚类结果进行评估,BHI 和 FOM 指标值分别见表 2 和表 3。

表 2 不同聚类集群数目下的 BHI 指标值

Tab.2 BHI index values under different number of clusters

算法类型\ 集群数目	CCA-Hc	EUC-Hc	差异率
$K=2$	0.466	0.233	100.05%
$K=4$	0.463	0.346	33.77%
$K=6$	0.467	0.377	23.90%
$K=7$	0.467	0.412	13.34%
$K=9$	0.465	0.435	7.12%
$K=11$	0.464	0.451	2.72%
$K=12$	0.463	0.456	1.48%

表 3 不同聚类集群数目下的 FOM 指标值

Tab.3 FOM index values under different number of clusters

算法类型\ 集群数目	CCA-Hc	EUC-Hc	差异率
$K=2$	2.697	4.633	-41.78%
$K=4$	2.697	4.298	-37.26%
$K=6$	2.696	4.047	-33.37%
$K=7$	2.696	3.995	-32.52%
$K=9$	2.696	3.816	-29.35%
$K=11$	2.695	3.693	-27.03%
$K=12$	2.695	3.636	-25.89%

表 2 中的差异率指的是 CCA-Hc 的 BHI 指标比 EUC-Hc 的 BHI 指标相差的百分比,同理可以计算表 3 中的差异率。

根据表 2 和表 3 的实验指标数据发现,对于 7 组不同的聚类集群数目实验,本文提出的 CCA-Hc 的 BHI 指标均高于 EUC-Hc, FOM 指标均低于 EUC-Hc,这表明 CCA-Hc 的鲁棒性更好,聚类结果中同一集群的基因相关性更大,聚类效果更加显著。同时还发现,集群数目对 CCA-Hc 的影响较小, K 选不同的值, BHI 指标值稳定在 0.463~0.467 之间, FOM

指标值稳定在 2.695 ~ 2.697 之间,而集群数目对 EUC-Hc 算法的影响相对比较明显。

图 2 为 CCA-Hc 在 *Oryza sativa* L. 数据集的聚类树状图,可以自行在所需的层级对树状图进行“剪枝”操作以获得合适的聚类效果^[21]。

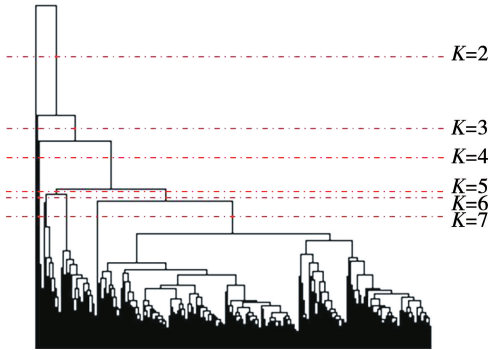


图 2 CCA-Hc 在 *Oryza sativa* L. 数据集的聚类树状图
Fjg.2 Clustering dendrogram of CCA-Hc in *Oryza sativa* L. dataset

3 结 论

本文为了充分有效地挖掘基因表达数据所蕴含的生物机制,提出一种基于 CCA – 层次聚类的基因聚类算法 (CCA-Hc)。把典型相关分析方法引入到凝聚层次聚类中进行多特征基因的聚类分析,成为本文的创新之处。该算法利用典型相关分析方法度量基因之间的相关性程度,能够充分考虑基因的多个特征信息。同时采用凝聚层次聚类可自主选择聚类集群数目,直观显示聚类结果。

基于 *Oryza sativa* L. (水稻)的基因表达数据集,本文对比了 CCA-Hc 和 EUC-Hc 的聚类效果,使用 BHI 和 FOM 两个评价指标进行衡量,结果表明 CCA-Hc 的鲁棒性和聚类准确性均更好,更有利于去探索基因表达数据潜在的生物机制。

参 考 文 献:

[1] 欧阳玉梅. 基因表达数据聚类分析技术及其软件工具 [J]. 生物信息学, 2010, 8(2): 104.
OUYANG Yumei. Gene Expression Data Cluster Analysis Technology and Software Tools [J]. Bioinformatics, 2010, 8(2): 104.
[2] 高华成. 基于数据降维框架的基因聚类算法[D]. 南京:南京邮电大学, 2021.

[3] 姚登举, 詹晓娟, 张晓晶. 一种加权 K – 均值基因聚类算法 [J]. 哈尔滨理工大学学报, 2017, 22(2): 112.
YAO Dengju, ZHAN Xiaojuan, ZHANG Xiaojing. A Weighted K-Means Gene Clustering Algorithm[J]. Journal of Harbin University of Science and Technology, 2017, 22(2): 112.
[4] 方匡南, 陈远星, 张庆昭, 等. 双向聚类方法综述 [J]. 数理统计与管理, 2020, 39(1): 22.
FANG Kuangnan, CHEN Yuanxing, ZHANG Qingzhao, et al. Review of Bidirectional Clustering Methods [J]. Journal of Applied Statistics and Management, 2020, 39(1): 22.
[5] 吴明阳, 张芮, 岳彩旭, 等. 应用 K-means 聚类算法划分曲面及实验验证 [J]. 哈尔滨理工大学学报, 2017(1): 54.
WU Mingyang, ZHANG Rui, YUE Caixu, et al. Application of K-means Clustering Algorithm for Surface Division and Experimental Verification[J]. Journal of Harbin University of Science and Technology, 2017(1): 54.
[6] LACKO D, HUYSMANS T, VLEUGELS J, et al. Product Sizing with 3D Anthropometry and K-medoids Clustering [J]. Computer-Aided Design, 2017; S0010448517301173.
[7] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: A New Data Clustering Algorithm and Its Applications [J]. Data Mining and Knowledge Discovery, 1997, 1(2): 141.
[8] FUSHIMI T, MORI R. High-Speed Clustering of Regional Photos Using Representative Photos of Different Regions [C]. 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018: 520.
[9] AI-MAMORY S O, KAMIL I S. A New Density Based Sampling to Enhance DBSCAN Clustering Algorithm [J]. Journal of Computer Science, 2019, 32(4): 315.
[10] ANKERST M, BREUNIG M M, KRIEDEL H P, et al. OPTICS: Ordering Points to Identify the Clustering Structure [C]// SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1 – 3, 1999, Philadelphia, Pennsylvania, USA. ACM, 1999: 2008, 99.
[11] 王飞, 王国胤, 李智星, 等. 一种基于网格的密度峰值聚类算法 [J]. 小型微型计算机系统, 2017(5): 1034.
WANG Fei, WANG Guoyin, LI Zhixing, et al. A Grid-based Density Peak Clustering Algorithm [J]. Journal of

- Chinese Computer Systems, 2017(5): 1034.
- [12] YAO J, CHANG C, SALMI M L, et al. Genome-scale Clusteranalysis of Replicated Microarrays Using Shrinkage Correlation Coefficient [J]. BMC Bioinformatics, 2008, 9: 288.
- [13] HONG S, CHEN X, JIN L, et al. Canonical Correlation Analysis for RNA-seq Co-expression Networks [J]. Nucleic Acids Res, 2013, 41(8): e95.
- [14] 万静, 郑龙君, 何云斌, 等. 高维数据的高密度子空间聚类算法 [J]. 哈尔滨理工大学学报, 2020, 25(4): 84.
- WAN Jing, ZHENG Longjun, HE Yunbin, et al. High-Density Subspace Clustering Algorithm for High-Dimensional Data [J]. Journal of Harbin University of Science and Technology, 2020, 25(4): 84.
- [15] 刘昊. 基于聚类算法的生物分析软件的设计与实现 [D]. 上海: 复旦大学, 2013.
- [16] 乔锦荣, 原新鹏, 梁旭东, 等. 凝聚层次聚类方法在降水预报评估中的应用 [J]. 干旱气象, 2022, 40(4): 690.
- QIAO Jinrong, YUAN Xinpeng, LIANG Xudong, et al. Application of Agglomerative Hierarchical Clustering Method in Precipitation Forecast Evaluation [J]. Arid Meteorology, 2022, 40(4): 690.
- [17] JASKOWIAK P A, CAMPELLO R J, COSTA I G. On the Selection of Appropriate Distances for Gene Expression Data Clustering [J]. BMC Bioinformatics, 2014, 15(2): 1.
- [18] 季姜帅, 裴颂文. 面向异质基因数据的智能层次聚类算法研究 [J]. 小型微型计算机系统, 2021, 43(9): 1808.
- Ji Jiangshuai, PEI Songwen. Research on Intelligent Hierarchical Clustering Algorithm for Heterogeneous Genetic Data [J]. Journal of Chinese Computer Systems, 2021, 43(9): 1808.
- [19] DATTA S, DATTA S. Methods for Evaluating Clustering Algorithms for Gene Expression Data Using a Reference Set of Functional Classes [J]. BMC Bioinformatics, 2006, 7(1): 1.
- [20] DATTA S. Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data [J]. Bioinformatics, 2003, 19(4): 459.
- [21] HULOT A, CHIQUET J, JAFFRÉZIC F, et al. Fast Tree Aggregation for Consensus Hierarchical Clustering [J]. BMC Bioinformatics, 2020, 21(1): 12.

(编辑: 温泽宇)