基因表达数据的聚类分析研究进展

岳 峰[1,2] 孙 亮[2] 王宽全[1] 王永吉[2] 左旺孟[1]

摘 要 基因表达数据的爆炸性增长迫切需求自动、有效的数据分析工具. 目前聚类分析已成为分析基因表达数据获取生物学信息的有力工具. 为了更好地挖掘基因表达数据, 近年来提出了许多改进的传统聚类算法和新聚类算法. 本文首先简单介绍了基因表达数据的获取和表示, 之后系统地介绍了近年来应用在基因表达数据分析中的聚类算法. 根据聚类目标的不同将算法分为基于基因的聚类、基于样本的聚类和两路聚类, 并对每类算法介绍了其生物学的含义及其难点, 详细讨论了各种算法的基本原理及优缺点. 最后总结了当前的基因表达数据的聚类分析方法, 并对发展趋势作了进一步的展望.

关键词 DNA 微阵列, 基因表达数据, 聚类分析
中图分类号 TP39

State-of-the-art of Cluster Analysis of Gene Expression Data

YUE Feng[1,2] SUN Liang[2] WANG Kuan-Quan[1] WANG Yong-Ji[2] ZUO Wang-Meng[1]

Abstract The flood of gene expression data provided by the DNA microarray technology has driven the development of automated analysis techniques and tools. Cluster analysis is an effective and practical method to mine the huge amount of gene expression data to gain important genetic and biological information. Many improved conventional clustering algorithms as well as new clustering algorithms have been proposed recently to process the gene expression data. This survey first introduces how to produce and represent the gene expression data, and then discusses the state-of-the-art cluster algorithms applied to gene expression data. According to the goals of clustering, clustering algorithms are divided into three categories: gene-based clustering, sample-based clustering, and biclustering. Basic biological principles and challenges for each category are presented. For each category, the basic principle is discussed in detail as well as its advantages and drawbacks. This paper concludes with a summarization in this field and a discussion of future trends.

Key words DNA microarray, gene expression data, cluster analysis

DNA 微阵列 (DNA microarray) 技术的迅速发展导致了基因表达数据 (Gene expression data) 的爆炸性增长, 同时大量的基因表达数据能够在公共数据库中得到[1]. 目前面临的主要问题是如何从大量的基因表达数据中利用自动分析工具得到有用的信息. 聚类分析 (Cluster analysis)[2] 作为一种有效的数据分析工具, 已广泛地应用于图像处理、信息检索、数据挖掘等领域. 作为一种探索性的分析工具, 聚类分析在基因表达数据的分析中已有广泛的应用, 主要包括: 通过对基因聚类发现未知基因的功能; 对样本聚类发现样本的显性结构 (Phenotype structure) 以及自动地对病理特征或实验条件进行分类; 通过两路聚类找出在某些条件下参与调控的基因聚类等. 在目前的生物信息学 (Bioinformatics) 研究中, 用于基因表达数据聚类分析的新算法、新软件包不断出现. 目前国内还没有关于基因表达数据的各种聚类算法的系统阐述. 本文探讨了基因表达数据聚类分析的原理、难点, 系统地总结了近年来提出的各种新算法, 并讨论了将来的发展趋势.

收稿日期 2006-11-01 收修改稿日期 2007-10-26
Received November 1, 2006; in revised form October 26, 2007
国家高技术研究发展计划 (863 计划) (2006AA01Z308) 和国家自然科学基金 (60373053, 60571025) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2006AA01Z308) and National Natural Science Foundation of China (60373053, 60571025)
1. 哈尔滨工业大学计算机学院生物信息技术研究中心 哈尔滨 150001
2. 中国科学院软件研究所互联网软件技术实验室 北京 100080
1. Biocomputing Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001
2. Laboratory for Internet Software Technologies, Institute of Software, Chinese Academy of Sciences, Beijing 100080
DOI: 10.3724/SP.J.1004.2008.00113

1 研究背景

1.1 基因表达数据的获得和表示

基因表达数据反映的是通过直接或间接测量得到的基因转录产物 mRNA 在细胞中的丰度 (Abundance), 目前基因表达数据主要通过 cDNA 微阵列 (cDNA microarrays)[3−5] 和寡核苷酸微阵列 (Oligonucleotide microarrays, 又称基因芯片、DNA 芯片)[6−8] 两种高通量检测技术获得. 它们的原理相同, 即利用四种核苷酸 (A, T, C, G) 之间两两配对互补的特性, 使两条在序列上互补的单核苷酸链形成双链, 该过程称为杂交 (Hybridization). 基本步骤是首先制备芯片, 在一个约 1 cm^2 大小的玻璃片上, 将称为探针 (Probe) 的 cDNA 或寡核苷酸片段固定在上面; 再从细胞或组织中提取

mRNA, 通过 RT-PCR 合成荧光标记的 cDNA, 与芯片上的 cDNA 或寡核苷酸片段杂交; 用激光显微镜或荧光显微镜检测杂交后的芯片, 获取荧光强度, 最后通过图像处理和分析得到细胞中 mRNA 丰度的信息.

在制造 cDNA 微阵列时, 采样点的大小不能保证完全一样, 因此不能通过直接比较不同微阵列图像的绝对荧光强度来比较 mRNA 的丰度. 通常使用双色荧光系统来纠正点之间的差异. 在制备样本时, 使用两个样本, 一个称为控制样本 (Control sample) 或对照样本 (Reference sample), 通常用绿色荧光素 (Cy3) 标记其 cDNA; 另一个称为测量样本, 用红色荧光素 (Cy5) 标记其 cDNA. 计算测量样本与对照样本之间荧光信号强度的比率或者对数化的比率, 从而比较一组实验中相对的基因表达水平. 在分析多个实验条件下的基因表达数据时, 对寡核苷酸芯片也采用一系列测量样本与对照样本之间的信号强度比率或比率的对数值作为实验结果.

下文没有特别说明时, 将 cDNA 微阵列和寡核苷酸芯片统称为 DNA 微阵列, 测得的数据称为基因表达数据, 将测量的对象统称为基因 (Gene), 将每次测量的条件称为样本 (Sample). 在对基因表达数据进行聚类分析时, 所得的数据都是在 n 种不同的条件 (样本) 下测得 m 个基因表达数据. 测量结果可以用 $m\times n$ 维矩阵 M 表示 (即 m 个基因, n 次测量), 用 x_{ij} 表示第 i 个基因在第 j 次实验中测得的值. 一般的, 对于在 n 次不同的测量中获得的关于同一基因的所有数据组成的向量称为该基因的基因表达向量, 如第 i 个基因的基因表达向量为: $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]$. 基因表达数据的特点是基因数目多, 样本少, 即 $m \gg n$. 通常 m 为数千甚至上万, 但是 n 较小, 一般为数十或者 100 左右. 虽然微阵列技术迅速发展, 但是一个组织的基因总数目的自然界限 (通常是 3 万到 10 万) 为数据集的规模确定了上限[9].

1.2 基因表达数据的预处理

在对数据进行聚类处理之前必须对原始数据进行预处理, 通常包括清除不完整的数据、合并重复数据、估计缺失数据 (Missing values)[10−12]、对数转化、删除在多次测量中变化不明显的基因, 规格化处理 (Normalization, 规格化之后每个基因对应的基因表达向量的各分量均值为 0, 方差为 1) 等[13]. 本文着重讨论聚类算法, 预处理不是讨论的重点.

1.3 聚类分析概述

聚类分析 (Cluster analysis) 即将待处理的对象分配到相应的聚类中, 使得同一聚类中的对象差别较小, 而不同聚类之间的对象差别较大. 与分类 (Classification) 不同的是, 聚类分析预先不知道类别信息, 没有训练集, 是一种非监督型 (Unsupervised) 的方法. 虽然其他领域如数据挖掘也有关于聚类分析的研究[2], 但由于基因表达数据自身的特点, 很多其他领域的算法并不完全适合处理基因表达数据.

目前, 在生物信息学领域提出了大量用于基因表达数据的聚类算法. 在基因表达数据分析中, 根据处理对象与目标的不同, 将聚类方法分为三类[14]: 基于基因的聚类 (Gene-based clustering)、基于样本的聚类 (Sample-based clustering) 和两路聚类 (Biclustering). 基于基因的聚类将基因看成聚类的对象, 将样本看成描述基因的特征. 在同一聚类中通常是表达模式类似的基因 (即共表达的基因, co-expressed gene), 它们一般具有相同的功能[15]. 这样利用聚类分析可由已知基因的功能推断同一聚类中未知基因的功能. 基于样本的聚类则以基因为特征, 以样本作为对象, 通过样本聚类, 可以发现样本的显性结构 (Phenotype structure), 自动对病理特征或实验条件进行分类[16]. 两路聚类 (Biclustering, 又称 subspace clustering, coclustering, direct clustering) 是指同时对基因和样本进行的聚类, 其目标是找出在某些条件下参与调控的基因聚类以及与某些基因相关联的条件, 从而更精确、更细致地探索基因和样本间的相互关系. 由于经典的聚类算法, 如 K 均值和层次聚类方法等, 已在相关的参考文献中作了较为全面的讨论, 本文不再赘述, 重点讨论近几年新出现的一些方法.

2 基于基因的聚类

2.1 意义及难点

基于基因的聚类以基因作为聚类的对象, 将样本作为基因的特征. 通过基因聚类, 可以发现表达模式类似的基因, 即共表达的基因. 由于在同一聚类的基因大都具有相同的功能, 因此可以根据聚类中已知基因的功能推断某些未知基因的功能[15].

在基于基因的聚类中, 由于噪声等原因, 存在大量的基因不属于任何一个聚类. 同时也存在一些基因在不同的条件下与不同的基因共表达, 即它们可属于多个聚类[17], 文献 [9] 中将这些基因称为 “intermediate gene”. 同时, 在聚类之后, 如何显示聚类结果, 如何利用生物学的背景知识来帮助分析, 都较难处理. 此外, 与数据挖掘中通用的聚类算法相比, 在对基因表达数据进行处理时更关注于算法的有效性 (Effectiveness), 即要求算法具有较高的准确性. 除了在数据分析问题中应用广泛的通用聚类方法, 近来, 学者们提出了更为复杂的也更加适宜于基

因表达数据分析的聚类方法, 包括基于人工神经网络的聚类算法和模糊聚类方法等, 下面分别加以介绍.

2.2 基于人工神经网络的聚类算法

在基于人工神经网络的聚类方法中, 自组织映射 (Self-organizing map, SOM)[18] 是目前使用较多的方法, 其核心思想是将 n 维向量通过合适的方式映射到一维或者二维空间中, 使人能够从视觉上更好地理解数据蕴涵的信息. 文献 [19] 在 SOM 的基础上提出了一种称为自组织树算法 (Self-organizing tree algorithm, SOTA) 的聚类方法, 结合了 SOM 和增长细胞结构 (Growing cell structures)[20] 两种方法的优点. 它与 SOM 最大的区别在于, 它使用了一种根据数据的具体特征不断改变拓扑形式的二叉树结构来刻画数据, 因此, 它能够得到数据的内在结构, 而不像 SOM 那样将数据局限于固定的拓扑形式中. SOTA 将聚类结果以层次结构的形式表现出来, 更加有利于聚类结果的分析, 而且具有更好的鲁棒性 (Robustness) 和较高的效率. 不足之处是 SOTA 生成的层次结构是二叉树, 这样就强制将数据表示为二叉树.

为了克服 SOTA 的缺点, 文献 [21] 提出了动态生长自组织树算法 (Dynamically growing self-organizing tree, DGSOT). DGSOT 与 SOTA 最大的不同在于, DGSOT 在建立每层的层次结构时自动选择最优的分支数目, 而不是强制设为 2. 在算法的实现上, 通过设计并交替使用垂直增长 (Vertical growth) 和水平增长 (Horizontal growth) 来对树进行扩张. 因此, 生成的层次结构不是二叉树, 而是根据具体数据分布的特征生成的多叉树. DGSOT 计算复杂度为 $O(m\log_d m)$, 其中 m 是待聚类的基因数, d 是生成的树的平均分支数. 相比传统的层次聚类算法, DGSOT 在效率上提高很大.

2.3 模糊聚类算法

很多蛋白质在细胞中有多种功能, 并且通过与不同的蛋白质合作来完成各自的功能. 这些蛋白质对应的基因与多组不同的基因在不同的条件下共表达, 每组基因被不同的调控机制所控制来适应细胞中变化的需求[17]. 换言之, 一个基因具有多面性, 在不同的条件下可能与不同的基因组共表达, 同时各个聚类之间有重叠的现象. 如从酵母基因表达数据中就可以推断出酵母基因中具有根据特殊条件而共表达的重叠的基因集合[17].

一般的非模糊聚类算法无法识别出与多组不同的基因表达模式相类似的基因, 难以获得在不同条件下与多组基因共调控的基因之间的关系. 特别是如果数据是由在多个实验条件下的数据合并而成时, 该缺陷更加显著. 而模糊聚类 (Fuzzy clustering) 算法则较好地解决了该问题. 模糊聚类算法最大的特点不是强制将每个基因归入到某个具体的聚类中, 而是计算每个基因对各个聚类的隶属度 (Membership). 用 $u_{ij}(0\leq u_{ij}\leq 1)$ 表示基因 i 对聚类 j 的隶属度, 且 u_{ij} 越大, 基因 i 隶属聚类 j 的程度越高. 最为经典的模糊聚类算法是模糊 C-均值 (Fuzzy C-means, FCM) 算法[22]. FCM 算法的思路与 K-均值算法类似, 不同的是在计算每个聚类的质心的时候要考虑隶属度, 同时每个基因不是直接归入某一个聚类. 设有 m 个基因及 k 个聚类, 首先初始化基因 i 与聚类 j 的隶属度 u_{ij} $(1\leq i\leq m, 1\leq j\leq k, 0\leq u_{ij}\leq 1)$, 得到隶属度矩阵 $U=[u_{ij}]_{m\times k}$; 然后使用公式 $\boldsymbol{c}_j=(\sum_{i=1}^m u_{ij}^q \boldsymbol{x}_i)/(\sum_{i=1}^m u_{ij}^q)$ 计算每个聚类的质心 $\boldsymbol{c}_j$, 这里 q 是模糊参数 (Fuzziness parameter), $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$ 是基因表达向量的集合; 再利用得到的质心计算每个基因到各个聚类的隶属度 $u_{ij}=\left(\sum_{s=1}^k \left(\frac{d(\boldsymbol{x}_i, \boldsymbol{c}_j)}{d(\boldsymbol{x}_i, \boldsymbol{c}_s)}\right)^{\frac{2}{q-1}}\right)^{-1}$. 当隶属度矩阵的变化较小时认为算法收敛, 结束迭代. FCM 算法的输出为若干个质心以及每个基因对各个聚类的隶属度值. 文献 [17, 23−24] 使用模糊聚类算法分析基因表达数据.

文献 [17] 采用的是 FCM 算法的一种变体, 成功识别出重叠的聚类, 揭示每个基因功能和调控 (Function & regulation) 的不同特征, 反映了酵母转录因子的目标和环境条件之间的关联. 同时构建了连续的聚类, 即对于每个聚类, 可将基因按隶属度的值从大到小进行排列, 这样得到所谓的 “连续的” 聚类. 此外, 相对于 K 均值算法, 算法对于聚类数目 k 不是太敏感, 稍大的 k 值不会严重影响算法的性能. 但是该算法也有若干缺陷: 1) 算法无法识别出所有的聚类, 在处理酵母基因表达数据时, 该算法识别出 90% 的已知聚类, 但是它无法识别出一些能够被层次聚类算法识别的聚类; 2) 参数太多, 难于设置, 如隶属度的阈值, 在很多情况下很难由用户手工确定最好的阈值; 3) 算法对数据进行 3 次 FCM 处理, 是根据酵母基因表达数据的特点设计的, 对于推广到其他数据集上则没有保证; 4) 对于数据较多和聚类数目较大时, FCM 算法的性能明显下降.

文献 [23] 指出在使用 FCM 算法对基因表达数据进行处理时, 通常将模糊参数设置为 2 的做法对于基因表达数据是不合适的, 原因有两个: 1) 计算得到的所有的隶属度值都相似, 即 $1/k$ (k 是聚类数目), 这样实质上没有提取出任何聚类的结构; 2) 虽然有时提出了聚类结构, 但所有的隶属度值都相对较低, 这意味着每个基因和每个聚类之间的联系都较弱. 文献 [23] 的关键贡献是提出了一种设置

模糊参数 q 的方法. 在 FCM 中, 如果 $q\to\infty$, 所有的隶属度 $u_{ij}\to 1/k$. 该文假设当 q 变化时, 隶属度矩阵 $U=[u_{ij}]_{m\times k}$ 和基因距离集合 Y_q 的变异系数 (Coefficient of variation) $cv(Y_q)$ 存在依赖关系. 通过实验数据发现, q 的变化导致 u_{ij} 趋向 $1/k$ 时, $cv(Y_q)$ 也趋向于 $0.03p$, 这里 p 是基因向量的维数. 由于导致 u_{ij} 趋向 $1/k$ 的 q 被认为是已趋于其上界值, 所以该文认为使得 $cv(Y_q)=0.03p$ 的值就是其上界 q_{ub}, 从而找到 q 的上界值 q_{ub}. 一般的, 在 FCM 中要求 $q>1$. 该文中将 q 设置在 1 和 2 之间, 并且尽量接近 2, 从而解决了该文提出的问题. 该文不足之处是关于模糊参数及其上界的公式都是经验公式, 缺乏相应的理论推导.

针对 FCM 算法中模糊参数难于设置的问题, 文献 [24−25] 从理论上给出了选取模糊参数的规则, 指出模糊参数的选择理论上依赖于数据本身. 令 $C_x=\sum_{k=1}^{n}\frac{(\boldsymbol{x}_k-\overline{\boldsymbol{x}})(\boldsymbol{x}_k-\overline{\boldsymbol{x}})^{\mathrm{T}}}{n\|\boldsymbol{x}_k-\overline{\boldsymbol{x}}\|^2}$, $\lambda_{\max}(C_x)$ 是矩阵 C_x 的最大特征值, 通过对目标函数的推导, 得到选取模糊参数的规则: 当 $\lambda_{\max}(C_x)<0.5$ 时, 需满足模糊参数 $q\le(-2\lambda_{\max}(C_x))^{-1}$, 而当 $\lambda_{\max}(C_x)\ge 0.5$ 时, 可以考虑根据经验选择 $q>1$ 的值[24]. 实验结果证明了该理论规则的有效性.

文献 [26] 采用了模糊 J-均值 (Fuzzy J-means, FJM) 算法和可变邻域搜索 (Variable neighborhood search, VNS)[27] 来处理基因表达数据. 在该算法中首先定义目标函数为 $R_q(V)=\sum_{i=1}^{m}\left(\sum_{j=1}^{k}\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^{2(1-q)}\right)^{(1-q)}$, 这里 $V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_k\}$ 是 k 个聚类质心组成的集合, $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$ 是基因表达向量的集合. FJM 的优化过程与 FCM 类似, 但是每次得到全部聚类质心以后, 去掉 V 中的一个 $\boldsymbol{v}_i$, 使得 $R_q(V/\{\boldsymbol{v}_i\})$ 最小. 再从与 $V/\{\boldsymbol{v}_i\}$ 关系不大 (或者与任意质心都不相似) 的基因中选择一个 $\boldsymbol{x}$, 使得 $R_q(V/\{\boldsymbol{v}_i\}\cup\{\boldsymbol{x}\})$ 最小. 将 V 更新为 $V/\{\boldsymbol{v}_i\}\cup\{\boldsymbol{x}\}$. 再根据 V 得到隶属度矩阵 $U=[u_{ij}]_{m\times k}$, 并更新 V(此处与 FCM 原理相同). 根据 V 计算目标函数 $R(V)$, 如果 $R(V)<R_{opt}$ 则将 R_{opt} 置为 $R(V)$, 并再对 V 进行删除 – 插入操作, 否则认为找到最优解. 文献 [26] 还引入了 VNS, 将 FJM 作为 VNS 内部循环的一个步骤. VNS 的思想类似于遗传算法 (Genetic algorithm), 通过随机变换 FCM 的初始输入, 将新的输入赋给 FJM 计算最优解. 一旦达到算法规定的资源限制 (如总的 CPU 时间, 总的循环次数等), 则算法结束, 并将当前的最优解置作为最终解. 总的来讲, VNS 在内部循环中使用 FJM 算法作为局部搜索工具, 使用类似于遗传算法来进行突变以克服局部搜索算法的缺陷.

将 VNS 和 FJM 相结合, 有效提高了算法的精度和性能. 实验中通过比较目标函数和 Jaccard 系数验证算法的聚类结果较好. 但是将 VNS 和 FJM 结合后计算复杂度较大, 此外, 设置模糊参数值的方法比较笨拙, 计算量较大.

与 K-均值类似, FCM 算法也只能得到超球形的聚类. 而文献 [28] 中提出了一种基于 Gustafson-Kessel (GK) 方法[29−30] 的模糊聚类算法, 可以找到不同几何形状的聚类. 通常距离定义为 $D_{xy}^2=\|\boldsymbol{x}-\boldsymbol{y}\|^2=(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}A(\boldsymbol{x}-\boldsymbol{y})$, 其中 A 是对称正定矩阵, 表示能够找到的聚类的几何形状. 当 $A=I$ 时就是常用的欧几里德距离, 它能找到的聚类是超球形的. 然而对所有的聚类都使用相同的 A 只能找到相同形状的聚类, 即使这样的聚类在实际数据中并不存在[28]. 因此, 作者提出基于自适应距离定义的 GK 方法, 对每个聚类使用不同的矩阵. 算法首先根据输入的聚类个数 k 随机确定 k 个质心和每个基因对各个聚类的模糊隶属度矩阵, 不断迭代, 使类内距离之和最小, 最终得到每个基因对各个聚类的模糊隶属度矩阵. 该算法的最大优点是能够找到不同几何形状的聚类, 能够更好地适应不同数据的特点. 缺点是需要估算的参数太多, 计算复杂性太高. 解决的办法是可以用其他快速的聚类算法 (如 K-均值) 的聚类结果作为该算法的初始值, 这样可以有效地减少迭代次数.

2.4 其他聚类方法

在基因表达数据的聚类分析中还有使用隐马尔可夫模型[31−32]、独立成分分析 (Independent component analysis, ICA)[33] 和模拟退火算法 (Simulated annealing)[34] 构建的算法. 当处理的基因表达数据是在较少的时间点测得时 (即如果数据是 m 个基因 $\times n$ 个时间点的规模, 且 $n\le 10$), 文献 [35] 提出了一种算法来专门处理此类短时间序列表达数据 (Short time series expression data).

3 基于样本的聚类

3.1 意义及难点

与基于基因的聚类不同, 基于样本的聚类是以基因为特征, 以样本作为聚类对象. 通过样本聚类, 可以发现样本的显性结构, 自动地对病理特征或实验条件进行分类. 此外, 更重要的是可以通过样本聚类找出与其相关的基因, 从而发现不同病理特征或实验条件下基因的调控机制.

由于在基因表达数据中基因维数要远远大于样本维数, 而在众多的基因中存在很多无关的或冗余的基因, 因此直接在基因表达数据上应用传统的聚类算法无法得到好的聚类结果. 为了降低信噪比, 有

必要选择一些与样本类别相关的基因 (称为信息基因, informative gene), 滤除对聚类无关或冗余的基因.

因此, 设计算法的关键在于找出信息基因, 即可以区分不同样本类别的基因. 由于聚类之前并不知道样本类别, 因此很难选出真正有区分度的基因, 而专家选择基因又没有通用性[36], 因此有必要通过其他方法自动地找出信息基因. 经典的基于样本的聚类方法有 CLIFF[37] 等, 此外, 近年来还出现了一些基于模型的样本聚类方法.

3.2 算法介绍

文献 [38−39] 讨论了将基于模型的算法应用于样本聚类的问题. 文献 [38] 中算法的原理基本上与文献 [40] 中基于模型的方法相同, 但是增加了利用主成分分析进行降维. 文献 [39] 中利用因子分析 (Factor analysis) 进行降维. 第一步, 选择与样本聚类相关联的基因. 方法如下: 假设样本集合是 k 个 T 分布组成的混合分布, 由于在对样本进行分类时, 一般聚类数目 $k = 2$ (如健康的和非健康的数据), 所以文献 [39] 中通过公式 $-2\log_2\lambda$ 来衡量基因对于样本聚类的重要程度, 这里 λ 是比较 $k = 1$ 和 $k = 2$ 的似然比统计量 (Likelihood ratio statistics). 同时选择满足如下条件的基因: 1) $-2\log_2\lambda > b_1$, 这里 b_1 是阈值; 2) $s_{\min} \geq b_2$, 这里 $s_{\min}$ 是该基因将样本分开之后的两个聚类中较小的聚类含有的样本数目, b_2 是阈值. 第二步分为 3 个子步骤: 1) 将所得的与样本关联较大的基因采用文献 [40] 中的算法聚类; 2) 利用每个聚类里的基因对样本进行聚类. 采用的技巧也是基于模型的, 称为因子模型的混合 (Mixtures of factor models)[41−42], 其中的降维方法是因子分析; 3) 计算每个基因聚类的平均向量, 利用它们对样本进行聚类. 这里 2) 和 3) 都得到了样本的聚类结果, 这时使用者可以加以分析, 并选取其中较好的结果.

文献 [43] 中提出了一种基于变化贝叶斯混合模型 (Variational Bayesian mixture model)[44−45] 的方法. 该方法假设数据为多层高斯分布的叠加, 根据变化贝叶斯混合模型定义损失函数, 不断进行迭代, 直到损失函数达到最小值, 结果收敛. 该算法可以自动得到聚类个数和各个聚类的参数, 每个样本可根据概率最大化方法进行分类, 这样就得到了样本的聚类结果. 在实际数据中, 本文使用 ICA[46−48] 得到二维数据, 再在二维数据上进行贝叶斯模型的估计. 与文献 [49] 中基于 EM+BIC 的方法相比, 该方法可以更加准确地发现数据的内部结构, 从而更加准确地聚类.

4 两路聚类

4.1 意义及难点

两路聚类是指同时对基因和样本进行的聚类. 由于在基因表达数据中: 1) 只有某些基因参与待考察的生物学过程; 2) 只有在某些条件下才会发生待考察的生物学过程[50], 因此, 两路聚类的目标是找出在某些条件下参与调控的基因聚类和与某些基因相关联的条件. 两路聚类可以看作一种 “局部的” 聚类: 由一部分基因确定样本集合或由一部分样本确定基因集合. 相比之下, 基于基因的聚类和基于样本的聚类都是 “全局的”: 以全部样本作为特征或以全部基因 (或选出的全部信息基因) 作为特征.

求解基因表达数据最优的两路聚类是 NP 问题. 因此, 现有的两路聚类算法大都是采用启发式方法求解目标聚类, 在聚类过程中对基因和样本同等看待. 在两路聚类的结果中, 基因或样本可以属于多个聚类, 也可以不属于任何聚类. Coupled two-way clustering (CTWC)[51−52], Biclustering[53] 和 plaid 模型[54] 等两路聚类方法已成功地应用在基因表达数据的分析中, 此外, 一类新的聚类算法 — 投影聚类算法也被应用到了基因表达数据上.

4.2 算法介绍

投影聚类 (Projected clustering) 首先由 Aggarwal 等[55] 提出, 指的是聚类对象 (Object) 在某些维特征上投影后再进行聚类. 由于在高维数据中, 可能很多维特征都是噪声, 真正的聚类只由全部特征的某个子集表示, 每个聚类都有一个投影的特征子集将其与其他对象区分开. 只有当某个特征可以将聚类中对象与其他对象区分开时, 这个特征才与此聚类相关. 因此投影聚类算法的输出不仅包括各个聚类所包含的对象, 还包括每个聚类对应的特征子集, 故可将投影聚类看作是一种两路聚类方法. 由于在基因表达数据中一组相关基因可能只在一部分样本中同时表达, 同样, 某些相关的样本可能只有一部分同时表达的基因, 因此投影聚类特别适合分析基因表达数据.

HARP (Hierarchical approach with automatic relevant attribute selection for projected clustering) 是由 Kevin 等[56−57] 提出的层次性投影聚类算法. 算法中定义每个属性对于聚类的相关指数 (Relevance index). 相关指数越大, 说明该属性越能够将其与其他聚类分离开. 合并指数 (Merge score) 用来衡量将两个聚类合并为一个聚类的优劣程度. 算法首先将每个对象都作为一个聚类, 再计算每两个聚类之间的合并指数, 合并具有最大合并指数的两个聚类, 并计算合并后聚类的各个属性的相

关指数. 对于超过给定阈值 $R_{\min}$ 的特征作为该聚类对应的特征子集. 重复上述合并聚类的过程, 直到只剩下一个聚类. 在选择合并后聚类对应的特征子集时, 只考虑属于两个合并前聚类的特征. 作者还使用了动态阈值放松 (Dynamic threshold loosening) 来避免使用预定义的参数: 在合并聚类的过程中, 只有大于 $R_{\min}$ 的特征才会被选择, 只有公共特征数目达到 $D_{\min}$ 的聚类才会被合并, 而 $R_{\min}$ 和 $D_{\min}$ 都随着聚类合并的过程由最大值向最小值线性递减. 在淋巴瘤的基因表达数据[58] 上应用 HARP 算法, 得到了若干个样本聚类, 每个样本聚类确定了一组具有生物学意义的基因, 其中一些是样本类型或生物学过程的指示基因[57].

该算法不需要事先输入聚类的质量和对应的特征子集的大小等参数, 又采用动态阈值放松避免预定义的阈值参数, 因此聚类结果对参数的设置不敏感. 但 HARP 也没有克服层次聚类算法的致命弱点: 易受噪声的影响. 如果作出了错误的决定将无法挽回, 而且会影响后续的聚类过程; 另一点就是强制性地将每个对象都聚到某一类中. 算法的时间复杂性较高 ($O(N^2d^2 + N^2\log^2(N))$), 其中 N 是对象个数, d 是特征维数, 使之不适合处理较大规模的数据.

5 总结和讨论

通过研究近几年出现的新方法可以看出, 基因表达数据的聚类算法已不再停留在早期的 K-均值、层次聚类等通用聚类方法上, 而是向着多元化、专门化、复杂化的方向发展. 同时, 更加适合基因表达数据特点的方法, 如模糊聚类方法和投影聚类方法受到越来越多的重视. 由于基因表达数据的规模不是很大, 因此相对于处理大量数据的聚类算法来说, 用于基因表达数据的聚类分析的方法可以有较大的时间及空间复杂性, 相应地, 也需要有更高的精度. 此外, 基因表达数据的聚类分析方法主要目标就是为生物学研究者提供一种探索基因表达数据的工具, 因此易用性及灵活性也是该类方法需要重点考虑的问题.

随着大量基因表达数据的产生, 越来越多的聚类算法被提出, 同时新的软件一般集成了多种算法. 实际使用中, 生物学研究者面临的一个重要问题就是如何选择合理的算法. 事实上, 并无绝对意义上的最优算法, 也没有公认的评价算法优劣的标准. 在实际使用中, 应该根据不同的用户需要以及数据特征选取不同的算法. 如果需要考察哪些基因在多个聚类中发挥作用时, 可以使用模糊聚类算法或者文献 [9] 中的交互型算法; 如果数据分布中聚类是呈现球形, 则 K-均值算法和 Adap_Cluster 算法[59] 都是很好的选择. 当然, 在使用算法时必须对算法的输入参数有较好的理解, 这样才能得到较好的结果. 其次, 在实际使用中可以将多种算法联合使用. 不同算法揭示了数据不同方面的特征, 从而可以加深对数据的理解. 如需要对数据有一个大致而直观的了解时, 可以使用层次聚类算法, 并可配合可视化工具, 如 TreeView.

聚类分析的最终目标就是帮助用户获取更多的生物学信息, 所以, 聚类算法的使用以及聚类结果的分析必须充分利用已知的生物学知识和一些辅助手段. 在这个过程中, 良好的人机交互可以显著提高效率. 特别是当处理的数据规模较大时, 很多算法的聚类结果较难理解. 目前通常采用可视化技术来帮助分析聚类结果. 第一种可视化表示方式是热量图 (Heat map) 和树图 (Dendrogram). 所谓的热量图将表示基因表达数据的矩阵 M 通过染色的形式表示出来. 树图是系统发生分析中常用的物种进化关系的表示方法, 在这里反映了基因表达向量之间的关系, 通过不同层次的剪枝, 可以得到不同的基因子集. 第二种可视化表示方式是点线图, 点线图比热量图更直观地表示基因的表达水平和不同基因在相同条件下的差异. 另一种方法是误差棒 (Error bar) 图, 即一个图中作出一个聚类的质心对应的点线图, 图中的误差条表示该聚类中基因在各个样本上变化的区间 (如标准差).

基因表达数据的大量涌现将继续推动聚类算法的研究. 一方面, 新的生物学场景将会不断出现, 进而提出新的问题, 从而推动算法研究的发展. 如两路聚类和模糊聚类在基因表达数据上的应用就是深入分析生物学场景后提出的. 另一方面, 对于传统算法不适宜处理基因表达数据的缺点而做的改进, 如文献 [60] 对于层次聚类算法的改进等. 总之, 新的聚类算法将吸收和集成尽可能多的生物学知识, 以便更好地分析基因表达数据, 从中挖掘出更多有价值的信息.

References

1 Brown P O, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 1999, **21**(1): 33−37

2 Jain A K, Murty M N, Flynn P J. Data clustering: a review. *ACM Computing Surveys*, 1999, **31**(3): 264−323

3 Schena M, Shalon D, Davis R W, Brown P O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 1999, **270**(5235): 467−470

4 Schena M, Scalon D, Heller R. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 1996, **93**(20): 10614−10619

5 Ramsay G. DNA chips: state-of-the art. *Nature Biotechnology*, 1998, **16**(1): 40−44

6 Lockhart D J, Dong H, Byrne M C, Follettie M T, Gallo M V, Chee M S. Expression monitoring by hybridization to

high-density oligonucleotide arrays. *Nature Biotechnology*, 1996, **14**(13): 1675−1680

7 Lipshutz R J, Fodor S P, Gingeras T R, Lockhart D J. High density synthetic oligonucleotide arrays. *Nature Genetics*, 1999, **21**(1): 20−24

8 Harrington C A, Rosenow C, Retief J. Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology*, 2000, **3**(3): 285−291

9 Jiang D X, Pei J, Zhang A D. An interactive approach to mining gene expression data. *IEEE Transactions on Knowledge and Data Engineering*, 2005, **17**(10): 1363−1378

10 Kim H, Golub G H, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 2005, **21**(2): 187−198

11 Tuikkala J, Elo L, Nevalainen O S, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, 2006, **22**(5): 566−572

12 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 2001, **17**(6): 520−525

13 Herrero J, Diaz-Uriarte R, Dopazo J. Gene expression data preprocessing. *Bioinformatics*, 2003, **19**(5): 655−656

14 Jiang D X, Tang C, Zhang A D. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 2004, **16**(11): 1370−1386

15 Eisen M B, Spellman P T, Brown P O, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 1998, **95**(25): 14863−14868

16 Golub T R, Slonim D K, Tamayo P, Huard C, Gassenbeek M, Mesirov J P. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, **286**(5439): 531−537

17 Gasch A P, Eisen M B. Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering. *Genome Biology*, 2002, **3**(11): 1−22

18 Gohonen T. *Self-organizing Maps*. Berlin: Springer-Verlag, 2001

19 Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 2001, **17**(2): 126−136

20 Fritzke B. Growing cell structures − a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 1994, **7**(9): 1441−1460

21 Luo F, Khan L, Bastan F, Yen I L, Zhou J Z. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. *Bioinformatics*, 2004, **20**(16): 2605−2617

22 Bezdek J C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell: Kluwer Academic Press, 1981

23 Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data. *Bioinformatics*, 2003, **19**(8): 973−980

24 Yu Jian. On the fuzziness index of the FCM algorithms. *Chinese Journal of Computers*, 2003, **26**(8): 968−973
(于剑. 论模糊 C 均值算法的模糊指标. 计算机学报, 2003, **26**(8): 968−973)

25 Yu J, Cheng Q S, Huang H K. Analysis of the weighting exponent in the FCM. *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, 2004, **34**(1): 634−639

26 Belacel N, Cuperlovic-Culf M, Laflamme M, Ouellette R. Fuzzy J-means and VNS methods for clustering genes from microarray data. *Bioinformatics*, 2004, **20**(11): 1690−1701

27 Belacel N, Hansen P, Mladenovic N. Fuzzy J-means: a new heuristic for fuzzy clustering. *Pattern Recognition*, 2002, **35**(10): 2193−2200

28 Kim D W, Lee K H, Lee D. Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics*, 2005, **21**(9): 1927−1934

29 Gustafson D E, Kessel W C. Fuzzy clustering with a fuzzy covariance matrix. In: Proceedings of the 17th Conference on Decision and Control Including Symposium on Adaptive Processes. San Diego, USA: IEEE, 1979. 761−766

30 Guthke R, Schmidt H W, Hahn D, Pfaff M. Gene expression data mining for functional genomics using fuzzy technology. In: Proceedings of Advances in Computational Intelligence and Learning: Methods and Applications. New York, USA: Springer, 2002. 475−487

31 Ji X L, Yuan Y, Li Y D, Sun Z R. HMMGEP: clustering gene expression data using hidden Markov models. *Bioinformatics*, 2004, **20**(11): 1799−1800

32 Schliep A, Schonhuth A, Steinhoff C. Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, 2003, **19**(1): i255−i263

33 Lee S I, Batzoglou S. Application of independent component analysis to microarrays. *Genome Biology*, 2003, **4**(11): 76

34 Hinneburg A, Keim D A. An effcient approach to clustering in large multimedia database with noise. In: Proceedings of the 4th International Conference on Knowledge Discovery in Databases. New York, USA: AAAI, 1998. 58−65

35 Smet F D, Mathys J, Marchal K, Thijs G, Moor B D, Moreau Y. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 2002, **18**(5): 735−746

36 Tang C, Zhang L, Ramanathan M, Zhang A D. Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering. Washington D. C., USA: IEEE, 2001. 41−48

37 Xing E P, Karp R M. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, 2001, **17**(1): 306−315

38 McLachlan G J, Bean R W, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 2002, **18**(3): 413−422

39 Debashis G. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 2002, **18**(2): 275−286

40 Yeung K Y, Fraley C, Murua A, Raftery A E, Ruzzo W L. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 2001, **17**(10): 977−987

41 McLachlan G J, Peel D. *Finite Mixture Models*. New York: Wiley, 2000

42 McLachlan G J, Peel D. Mixtures of factor analyzers. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 2000. 599−606

43 Teschendorff A E, Wang Y Z, Barbosa-Morais N L, Brenton J D, Caldas C. A variational Bayesian mixture modeling framework for cluster analysis of gene-expression data. *Bioinformatics*, 2005, **21**(13): 3025−3033

44 Attias H. Inferring parameters and structure of latent variable models by variational Bayes. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 1999. 21−30

45 Wang B, Titterington M. Convergence Properties of a General Algorithm for Calculating Variational Bayesian Estimates for a Normal Mixture Model, Technical Report No. 04-2, Department of Statistics, University of Glasgow, UK, 2004

46 HyvÄarinen A, Karhunen J, Oja E. *Independent Component Analysis*. New York: Wiley, 2001

47 Martoglio A M, Miskin J W, Smith S K, MacKay D J C. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, 2002, **18**(12): 1617−1624

48 Liebermeister W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 2002, **18**(1): 51−60

49 Fraley C, Raftery A E. MClUST: software for model-based cluster analysis. *Journal of Classification*, 1999, **16**(2): 297−306

50 Madeira S C, Oliveira A L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2004, **1**(1): 24−45

51 Getz G, Gal H, Kela I, Notterman D A, Domany E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, 2003, **19**(9): 1079−1089

52 Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **97**(22): 12079−12084

53 Cheng Y Z, Church G M. Biclustering of expression data. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. Menlo Park, USA: AAAI, 2000. 93−103

54 Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica Sinica*, 2002, **12**(1): 61−86

55 Aggarwal C C, Procopiuc C M, Wolf J L, Philip S Y, Park J S. Fast algorithms for projected clustering. In: Proceedings of ACM International Conference on Management of Data. New York, USA: ACM, 1999. 61−72

56 Yip K Y. HARP: A Practical Projected Clustering Algorithm for Mining Gene Expression Data [Master dissertation], The University of Hong Kong, 2004

57 Yip K Y, Cheung D W, Ng M K. HARP: a practical projected clustering algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 2004, **16**(11): 1387−1397

58 Alizadeh A A, Eisen M B, Davis R E, Ma C, Lossos I S, Rosenwald A. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 2000, **403**(6769): 503−511

59 Smet F D, Mathys J, Marchal K, Thijs G, Moor B D, Moreau Y. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 2002, **18**(5): 735−746

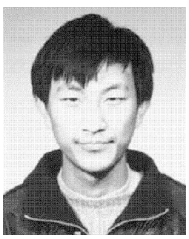
60 Bar-Joseph Z, Demaine E D, Gifford D K, Srebro N, Hamel A M, Jaakkola T S. K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, 2003, **19**(9): 1070−1078

岳 峰 哈尔滨工业大学计算机学院博士研究生. 主要研究方向为数据挖掘、生物识别、生物信息学. 本文通信作者.
E-mail: csfyue@gmail.com
(**YUE Feng** Ph. D. candidate at School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers data mining, biometrics, and bioinformatics. Corresponding author of this paper.)

孙 亮 中国科学院软件所硕士研究生. 主要研究方向为数据挖掘和生物信息学.
E-mail: sun.liang@asu.edu
(**SUN Liang** Master student at Institute of Software, Chinese Academy of Sciences. His research interest covers data mining and bioinformatics.)

王宽全 哈尔滨工业大学计算机学院教授, IEEE 高级会员. 主要研究方向为生物识别、生物计算、生物系统建模与仿真技术.
E-mail: wangkq@hit.edu.cn
(**WANG Kuan-Quan** Professor at School of Computer Science and Technology, Harbin Institute of Technology, senior member of IEEE. His research interest covers biometrics, bio-computing, and modeling and simulating biology system.)

王永吉 中国科学院软件所研究员. 主要研究方向为实时系统、人工智能、数据挖掘、软件工程.
E-mail: ywang@itechs.iscas.ac.cn
(**WANG Yong-Ji** Professor at Institute of Software, Chinese Academy of Sciences. His research interest covers real-time systems, artificial intelligence, data mining, and software engineering.)



左旺孟 哈尔滨工业大学计算机学院讲师. 主要研究方向为生物特征识别、生物信息学和生物系统建模与仿真技术.
E-mail: cswmzuo@gmail.com
(**ZUO Wang-Meng** Lecturer at School of Computer Science and Technology, Harbin Institute of Technology. His research interest covers biometrics, bioinformatics, and biosystem modeling and simulation.)