第 26 卷 第 1 期 华侨大学学报(自然科学版) Vol. 26 No. 1
2005 年 1 月 Journal of Huaqiao University (Natural Science) Jan. 2005

文章编号 1000-5013(2005)01-0007-04

聚类算法在基因表达数据分析中的应用

朱 婵 许龙飞

(暨南大学信息科技学院,广东 广州 510632)

摘要 聚类算法在基因表达数据的分析处理中得到日益广泛的应用. 文中对几种典型的聚类算法进行描述,对各算法在基因表达数据处理中的特点,进行评价并提出改进的策略. 最后,指出聚类算法在生物信息学应用中的发展趋势.

关键词 生物信息学,基因表达数据,聚类算法

中图分类号 Q 786 **文献标识码** A

生命科学与信息科学是目前发展最为迅速的两大领域. 生物信息学作为这两大学科交叉的产物之一,已在基因组学研究中发挥巨大的作用. 而凝结这两大学科研究成果的另一项崭新技术——基因芯片,已成为大规模提取和探索生物分子信息的强有力手段,将在后基因组研究中发挥突出的作用[1]. DNA 微阵列实验所产生的大量复杂数据给生物信息学研究者带来了严峻的挑战,需要使用准确可靠的工具对这些数据进行分析. 聚类方法是在基因组学研究领域应用最广泛的技术之一,其目标是将微阵列实验中表现出相似表达模式的基因聚集到一个簇中. 目前已有多项研究证明,聚类技术对于共调控基因的发现是卓有成效的[2,3]. 本文对基因表达数据分析中的聚类算法进行了分析和研究. 我们相应简要地介绍几种典型的聚类算法,对各聚类算法进行评价. 最后,指出聚类算法在基因表达数据分析中的发展和研究趋向.

1 常用的基因表达数据聚类算法

目前对基因表达数据的处理主要是进行聚类分析. 聚类是将数据对象分组成为多个类或簇,在同一个簇中的对象之间具有较高的相似度,而不同簇中的对象则差别较大. 通常采用数据对象属性值间的距离来进行相似或不相似的度量. 聚类分析通过各种不同的数学模型,对具有相同统计行为的多个基因进行归类,将表达规律相似的基因聚为一类. 在此基础上,寻找相关基因并分析基因的功能. 已有多种聚类算法,应用于对基因表达和蛋白质序列等进行聚类. 以下简要介绍最常用的简单聚类、层次聚类、K-平均聚类和自组织特征映射.

1.1 简单聚类

假设有 n 个基因的表达数据向量分别为 X_1, X_2, …, X_n, T 为预先定义的阈值. 下述简单聚类的算法过程. (1) 令任意一个基因的表达向量为第一个聚类的中心. (2) 依次处理其它基因. (a) 在处理第 i 个基因时,首先计算该基因的表达数据向量与现有各类中心的距离. (b) 假设其与第 j 类的距离 D_{ij} 最小,且 $D_{ij} < T$,则将基因 i 分配到第 j 类;否则,生成一个新类,且其中心为第 i 个基因的表达向量.

1.2 层次聚类

层次聚类先将每个对象作为单独的一个簇,然后相继合并距离最近的对象或聚类,最后的聚类结果

收稿日期 2004-04-19

作者简介 朱 婵(1979-),女,硕士研究生,主要从事信息系统与知识工程的研究. 许龙飞(联系人),男,教授,通信地址:510632 暨南大学羊城苑 10 栋 404. E-mail: txlf@jnu.edu.cn

基金项目 国家自然科学基金资助项目(60374070);广东省自然科学基金资助项目(031903)

由系统树图表示. 每个分枝都代表一个有相似行为的基因组. 聚类过程主要有3点. (1) 计算任意两个对象之间的距离,将距离最小的两个结合在一起,并生成一个新节点. (2) 计算剩下的对象中同新节点距离最近的,将其与新节点合并. (3) 迭代进行距离计算与合并的操作. 对于 n 个对象的情况,该过程最多重复 $n-1$ 次,直至最后剩下一个节点.

1.3 K-平均聚类

K-平均聚类是最典型的基于划分的聚类方法. 现作算法过程简单描述. (1) 任意选取 K 个基因表达向量作为初始聚类中心 Z_1, Z_2, …, Z_k. (2) 计算其它序列与各聚类中心的距离,将其划分到与之距离最小的聚类. (3) 重新计算新得到聚类的中心对象. (4) 对于所有的聚类中心,如果

$$Z_j(t+1) = Z_j(t), \qquad j = 1,2,\cdots,K,$$

则迭代结束,得到最后的聚类结果;否则,返回(2)继续迭代计算.

1.4 自组织特征映射

自组织特征映射[4]属于神经网络聚类方法的一种,能够将任意高维的输入转换到低维(二维)输出. 假定输入的基因表达为 N 维,输出节点数为 M,设信号按某种顺序输入. 即

$$x(t) = \{x_i(t), i = 1, \cdots: N\}.$$

下述算法过程. (1) 建立一个有 M 个输出节点的二维网格,对各节点的权值(N 维向量)用随机数初始化. (2) 在样本集中随机选某样本 $x(t)$ 输入. (3) 在输出节点集中寻找权向量与 $x(t)$ 最近的节点 y. (4) 与 y 在同一邻域的节点根据学习步长 $\mu(t)$ 和邻域大小 $h(t)$ 相应调整权值,其中 $\mu(t)$ 和 $h(t)$ 都随迭代次数而变化. (5) $t \leftarrow t+1$ 转向(2),直到邻域大小收缩到只包含中心本身.

2 算法的评价和改进

聚类分析是一个富有挑战性的研究领域. 由于生物数据的海量性、复杂性等特点,对聚类分析算法也提出特别的要求. 它包括比较典型的有处理高维数据的能力、处理不同类型数据的能力、发现具有任意形状的聚类的能力,以及聚类结果的可解释性等.

2.1 聚类算法的评价

前面所介绍的几种常用聚类算法,其中 K-平均算法的计算复杂度为 $O(nkt)$. 此处, n, k, t 分别为样本数、类别数和迭代次数,通常情况下, $k \ll n$, $t \ll n$. 因此, K-平均聚类可应用于数据量较大的情况,这是其优点之一. 算法的关键问题是如何初始化质心. 由于有多种初始化的可能,因此难以得到最优化的结果,且该算法只适合形状为凸形的聚类,而在生物中某些复杂形状的聚类则未必相符. 此外,样本向量各维的重要性也未必相同,算法未加考虑可能需要通过加权值处理. 由于采用一个类中所有对象的平均值作为质心,聚类结果还易受孤立点的影响. 自组织特征映射应用类间的全局关系,保证了拓扑有序性,使得输入数据中具有相似特征的点映射后在空间上也是邻近的. 它能够以少量聚类中心表示原数据,起到数据压缩的作用. 自组织特征映射能够为基因微阵列数据的可视化和分析处理提供高效的模式,但还不能直接用于分类或识别,需要对它再做一些监督学习. 层次聚类方法和基于划分的方法区别在于,它并不是试图寻找最佳的聚类结果,而是按照一定的相似性判别标准,对最相似的部分进行合并. 层次聚类方法简单直接,易于理解和应用. 但它适用于反映真正的层次树结构,而微阵列数据的产生往往并非如此. 其聚类结果受各个类的大小和其中对象分布形状的影响,适用于类的大小相似且对象分布为球形的聚类.

2.2 算法的改进策略

目前的基因表达微阵列实验尚处于早期阶段,多数用来确定聚类数目的方法都是非监督的. 因此,所得到的聚类未必与具有生物学意义的簇相对应. 聚类结果应该包含更多的生物学信息,而不是仅仅将数据点严格的划分成簇. 如何有效提高聚类结果的生物学意义,是尤为需要改进的一个方面. 由此提出将聚类数的有监督学习融入非监督的聚类算法,从而得到与生物学意义更相符的聚类数目[5]. 通过有监督学习得到的聚类数 k^*,定义为将所有在生物学意义上相关的基因分类到各个簇中之后得到的最大聚类数目为

$$k^{*} = \max\{k : \forall_{m' \in \{1,2,\cdots,k\}, m \in \{1,2,\cdots,k\}} \max\#(C_m \cap B_{m'}) \geq \beta'_m)(\#B_{m'})\}.$$

这里, C_m 为第 m 个聚类中的基因对应的索引集合, $B_{m'}$ 是所有已知属于第 m' 个已知生物学的簇的基因索引集合, k 是已确定的生物聚类的数目, β'_m 是一个基于生物学意义选择出的比例值, 且 $\beta'_m \in [0,1]$, # 表示其后集合中元素的个数. 这个式子表示, 当有一个满足给定约束条件的更大聚类数时, 该数即为符合生物学意义的最大聚类数目. 给定的约束能够更好地对基因进行区分, 并避免将已知在生物上相关的基因进行硬性的划分. 对自组织特征映射也可加以监督学习来改善聚类性能. 例如, 对聚类中心加以监督学习, 就可以将该中心代表的所有数据都归入中心的所属类别. 为了提高识别的正确率, 还可用训练样本对权值作"细调". Kohonen 提出学习向量量化算法(LVQ)[6], 使得聚类中心的权值更靠近被正确聚类的数据对象; 否则远离. 有兴趣的读者可查阅相关资料.

3　聚类算法的研究趋向

在目前的研究中, 已有多种聚类算法和大量的聚类程序被用于 DNA 微阵列数据的分析. 在实际应用中, 我们需要根据应用所涉及的数据类型、聚类的目的, 以及具体的应用要求, 选择合适的聚类算法, 并不断研究新的方法.

3.1　聚类算法间的结合

多数聚类算法都需要由用户决定聚类的数目. 但对许多基因表达数据, 我们往往无法确定数据所应划分出的聚类数目, 而更希望这个参数由被聚类的数据驱动来产生. 二叉层次聚类算法将层次聚类和 K-平均算法相结合来产生聚类数目的算法, 从而达到了完全的无监督. 下述该算法的具体步骤. (1) 使用 K-平均算法, 将表达数据划分为两个聚类. 从而, 得到聚类的质心(Centroid)和类的数目. (2) 计算(1)中所得两个聚类的 Fisher 线性判别, 若结果大于一个已经设定好的阈值(Threshold), 则接受此次聚类结果; 否则, 不执行该次划分. (3) 返回步骤(1), 对各聚类迭代运行此算法, 直到不再发生变化. 在该算法中, K-平均算法迭代运行, 不断更新各个聚类的质心和类成员. 其最大的改进, 在于无需预先设定类的分布状态和聚类的数目. 虽然在使用 Fisher 线性判别时也要设定一个阈值, 但是这个值是与聚类的空间结构相关的, 可以通过一些实验进行估计. 在仅仅使用 K-平均算法时, 要在没有任何先验知识的情况下得出聚类数目, 是非常困难的.

3.2　聚类算法与其它数据挖掘方法的结合

将聚类算法与其它算法相结合也是较为常用的策略. 例如, 先使用层次聚类算法对给定的基因表达数据集进行分类, 然后再根据基因的文本描述来消除某些聚类边界, 以使各聚类中基因表达的功能相关性达到最优化[7]. 或者, 针对许多聚类算法(例如 K-平均算法)需要事先给定聚类的数量的特点, 通过对聚类中数据点的密度进行对照来找出一个最优的聚类数量[8]. 还可将支撑向量机与序列比对算法相结合, 对具有较远亲缘关系的蛋白质序列进行同源探测[9].

3.3　聚类算法与领域知识相结合

目前聚类分析方法只能找出基因之间简单的、线性的关系, 主要基于统计学的理论而很少利用到生物领域的知识. 这既使结果由于缺乏领域内知识的约束而可能出现不合理性, 同时又失去了利用领域内知识优化算法的优势. 信息融合法[10~11]将领域知识和聚类分析方法相结合, 通过利用尽可能多的可用数据来探测新的信息. 该方法将用到的每一个数据集都看作一个类, 如基因表达数据、以 DNA 模体形式出现的序列信息、定位信息, 以及基因本体信息等. 通过同时运用不同类型的数据来发现有意义的模式, 从而对数据产生更深刻的理解.

3.4　加强聚类算法在分布式环境下的应用及其可视化表示

(1) 随着生物数据量的迅猛增长, 生物数据库中存储了大量的数据. 在这样的数据环境下进行聚类分析可能要占用大量的处理器资源, 需要用分布式系统来解决这一问题. 同时, 许多大型数据库本来就是分布式的, 这使得研究进行聚类分析的高效分布式算法非常重要. (2) 聚类算法的可视化表示. 将基因表达数据的聚类结果以图形方式显示出来, 而且能够和相关的数据库链接. 通过对表达数据的整合, 使用户能够直观地了解各类表达数据在序列水平上的信息.

4 结束语

在对聚类算法进行研究时,特别是在基因表达条件下,有很多重要问题需要考虑.例如,被聚类的数据类型的表示、相似度的度量方法、类别数的选取等.通过对相关基因的表达、突变、重排的分析研究,确定相应的模型与算法,发现具有协同调节和功能相关的组,对于完成某一生物体的全套基因有非常显著的意义.随着多个基因组测序的完成、微阵列技术的广泛应用,以及基因表达数据的大量积累,对聚类算法的敏感性、可伸缩性和运行速度都提出了更高的要求.此外,对各种聚类算法的有效性进行分析也是一项重要而艰巨的任务.

参 考 文 献

1 孙 啸,王 晔,何农跃等.生物信息学在基因芯片中的应用[J].生物物理学报,2001,17(1):27～34

2 Eisen M, Spellman P, Brown P, et al. Cluster analysis and display of genome-wide expression patterns[J]. Proc. Natl. Acad Sci. US,1998,(95):14 683～14 688

3 王富刚,陈先农.基因芯片数据的聚类分析[J].国外医学生物医学工程分册,2004,27(2):98～101

4 Kohonen T. Self-organizing maps[M]. 3rd ed. Heidelherg: Springer, 2001. 25～40

5 Bickel D R. Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically[J]. Bioinformatics,2003,19(7):818～824

6 Somervuo P, Kohonen T. Self-organizing maps and learning vector quantization for feature sequences[J]. Neural Processing Letters, 1999,10(2):151～159

7 Raychaudhuri S,Chang J T,Imam F,et al. The computational analysis of scientific literature to define and recognize gene expression clusters[J]. Nucleic Acids Research, 2003,15: 4 553～4 560

8 Wicker N,Dembele D,Raffelsberger W,et al. Density of points clustering,application to transcriptomic data anaylsis[J]. Nucleic Acids Research, 2002,(18):3 992～4 000

9 Li Liao,Noble W S. Combining pairwise sequence similarity and support vector machines for remote protein homology detection[J]. Recomb, 2002, (18):255～232

10 Kalton A,Wagstaff K,Yoo J. Generalized clustering,supervised learning, and data assignment [J]. KDD, 2001,(1):299～304

11 Baldi P, Brunak S著. 生物信息学——机器学习方法[M]. 张东晖等译. 北京:中信出版社,2003.276～279

Application of Clustering Algorithms to the Analysis of Gene Expression Data

Zhu Chan Xu Longfei

(College of Information Science and Technology, Jinan University, 510632, Guangzhou, China)

Abstract Clustering algorithms have become increasingly important in analyzing and processing gene expression data. Several typical clustering methods are described here. After estimating the characteristic of clustering methods in processing gene expression data, some strategies for its improvement are proposed; and the trend of applying clustering algorithms to bioinformatics is pointed out.

Keywords bioinformatics, gene expression, data, clustering algorithm