文章编号: 1001- 2486(2010) 04- 0150- 07

一种基于线性流形的基因表达数据的聚类方法*

黎刚果, 王正志, 王广云, 倪青山, 强 波
(国防科技大学 机电工程与自动化学院, 湖南 长沙 410073)

摘 要: 由于基因表达数据的稀疏性和噪声性, 传统聚类算法对其聚类时不能取得好的效果。针对这一问题, 一种新的线性流形方法被提出, 它的基本思想是搜索数据集中的线流形聚类, 再将其中某些线流形聚类融合构造高维流形聚类。该算法将切向距离和法向距离作为线性流形的距离度量, 运用空间近邻信息, 采用聚类基因的平均表达水平作为转移向量, 提高了聚类的准确度。实验结果表明, 该算法的聚类准确性优于其它聚类算法, 并且对带有噪声的数据可以保持较高的聚类准确度; 在对 Hela 基因表达数据聚类时, 算法得到了具有显著生物学意义的聚类。这些都说明提出的算法对基因表达数据聚类的适用性和有效性。

关键词: 基因表达数据; 线性流形; 子空间聚类; 线流形
中图分类号: Q- 332 文献标识码: A

A Clustering Method for Gene Expression Data Based on Linear Manifold

LI Gang-guo, WANG Zheng-zhi, WANG Guang-yun, NI Qing-shan, QIANG Bo
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Conventional clustering methods fail to obtain good clustering performances for gene expression data due to the inherent sparsity of data and the existence of noise. A new linear manifold clustering method was proposed to address this problem. The basic idea of this method is to search the line manifold clusters hidden in datasets and then fuse some of the line manifold clusters to construct higher dimensional manifold clusters. The method considers the orthogonal distance and the tangent distance as the linear manifold distance metrics, utilizes spatial neighbor information and takes the real gene expression profile as the transition vector. The experimental results show the superiority of this method over other competing clustering methods in terms of clustering accuracy and the anti-noise capability of this method. Moreover, the proposed method is able to obtain some clusters with significant biological meaning for Hela gene expression data. All these demonstrate the method proposed is suitable and valid for the gene expression data clustering.

Key words: gene expresion data; linear manifold; subspace clustering; line manifold

随着过去几年基因芯片技术的快速发展, 产生了大量的基因芯片实验数据(基因表达数据)。很多传统的聚类方法已经被用来对基因表达数据进行分析。然而传统的聚类方法用于基因表达数据聚类分析时不能取得很好的效果。因此出现了很多针对高维数据聚类的算法, 比如: CLIQUE[1]、PROCLUS[2]、ORCLUS[3] 等子空间聚类方法。因为 CLIQUE、PROCLUS 和 ORCLUS 方法都使用物理距离来度量数据点间的距离, 所以它们可能丢失数据间的相关性信息。而正相关、负相关和其它更复杂的相关性都能泛化为线性流形[4]。Haralick 和 Harpaz 提出了一种线性流形聚类方法 LMCLUS[5], 并将其运用到基因表达数据聚类中。然而他们的方法也存在缺陷。LMCLUS 需要估计数据中所蕴含的线性流形的最高维度, 但准确地估计线性流形的最高维度是很困难的。本文提出了一种新的基于线性流形的基因表达数据的聚类方法(line searching and fusing cluster, LSAFCLUS), 该方法的基本思想是搜索数据中的线流形聚类, 然后通过融合线流形聚类构造高维线性流形聚类。

* 收稿日期: 2010- 01- 05
基金项目: 国家自然科学基金资助项目(60835005)
作者简介: 黎刚果(1982—), 男, 博士生。

1 线性流形聚类

1.1 线性流形

定义 1(线性流形) 如果向量空间 V 中存在子空间 S 及转移向量 t 满足: $L=\{x\in V|\ s\in S,\ x=t+s\}$, 则 L 为向量空间上的线性流形, 其维数为子空间 S 的维度。如果 L 比向量空间 V 少一维, 则称 L 为 V 上的超平面。如果存在边界向量 a_L 和 a_H 使得 $L=\{x\in V|\ s\in S$, 且 s 在边界向量 a_L 和 a_H 所围区域之中, $x=t+s\}$ 成立, 则称 L 为矩形边界线性流形, 且其中心为 $t+\frac{a_L+a_H}{2}$。当 $a_L=-a_H$ 时, 其中心即为转移向量 t。一个线性流形可以通过一个对初始子空间进行转换后的线性子空间来表示。在 $\boldsymbol{R}^n$ 中, 一个维数为 $r(1\leqslant r\leqslant n)$ 的线性流形 L 的模型可表示为: $x=t+b_1a_1+\cdots+b_ra_r$。其中, x 是 L 中任一元素, t 是转移向量, $a_1,\ a_2,\ \cdots,\ a_r$ 是线性无关的向量, $b_1,\ b_2,\ \cdots,\ b_r$ 为参数。当线性流形的维数为 0 时, 流形表示为点; 当其维数为 1 时, 流形表示为线, 称之为线流形。在 $\boldsymbol{R}^2$ 中, 线性流形可以是点、线及其本身。如果线性流形 L_1 中的所有元素都属于流形 L, 则流形 L_1 为 L 的子流形。

定理 1 设 L 是 $\boldsymbol{R}^n$ 中一个 r 维 $(1\leqslant r\leqslant n)$ 的线性流形, 其模型为 $x=t+b_1a_1+\cdots+b_ra_r$($a_1,\ a_2,\ \cdots,\ a_r$ 是线性无关的向量, $b_1,\ b_2,\ \cdots,\ b_r$ 为参数) , 且设 p 为 L 中的一个元素, 则 L 的模型也可以表示为: $x=p+b'_1a_1+\cdots+b'_ra_r$。

证明: 既然 p 是 L 中的一个元素, 则 p 可表示为 $p=tp_1a_1+\cdots+p_ra_r$。移项可得 $t=p-p_1a_1-\cdots-p_ra_r$, 则 L 的模型可表示为: $x=p+(b_1-p_1)a_1+\cdots+(b_r-p_r)a_r$。令 $b'_1=b_1-p_1,\ \cdots,\ b'_r=b_r-p_r$, 则 L 的模型表示如下:

$$x=p+b'_1a_1+\cdots+b'_ra_r$$

定理 2 设 L_1 和 L_2 分别为 q 维和 $r-q(q<r)$ 维的线性流形, 且其模型分别表示为: $x=u+b_1a_1+\cdots+b_qa_q;\ y=t+b_{q+1}a_{q+1}+\cdots+b_ra_r$($a_1,\ a_2,\ \cdots,\ a_r$ 是线性无关的向量, $b_1,\ \cdots,\ b_r$ 是参数。)。L 为 r 维线性流形, 其模型表示为: $z=w+c'_1a_1+\cdots+c'_ra_r$($c_1,\ \cdots,\ c_r$ 是参数)。如果元素 $w\in L_1$ 且 $w\in L_2$, 那么 L_1 和 L_2 是流形 L 的子流形。

证明: 因为元素 $w\in L_1$ 且 $w\in L_2$, 所以由定理 1 可知, 它们的模型可分别表示为: $x=w+b'_1a_1+\cdots+b'_qa_q;\ y=w+b'_{q+1}a_{q+1}+\cdots+b'_ra_r$。既然 L 的模型可表示为: $z=w+c'_1a_1+\cdots+c'_ra_r$, 那么当 $c'_1,\ c'_2,\ \cdots,\ c'_q$ 全为 0 时, 流形 L 成为 L_1, 而当 $c'_{q+1},\ c'_{q+2},\ \cdots,\ c'_r$ 全为 0 时, 流形 L 成为 L_2。因此, L_1 和 L_2 是流形 L 的子流形。

由定理 2 可知, 当两个线流形满足定理 2 的条件时, 则可以构造出一个二维流形。同时定理 2 只描述了两个流形的情况, 可以扩展到多个流形的情形。其证明过程与定理 2 的证明类似。

定理 2 的补充 定理 2 中要求 $a_1,\ a_2,\ \cdots,\ a_r$ 是线性无关的向量。当此条件不满足时, 不妨设 $a_1,\ \cdots a_m,\ a_{q+1},\ \cdots,\ a_{q+n}$ 为向量 $a_1,\ a_2,\ \cdots,\ a_r$ 中的一个极大线性无关组($m+n=p<r$, $1\leqslant m\leqslant p$, 且 $1\leqslant n\leqslant r-q$) , 且设一个 p 维流形 L 的模型表示为: $z=w+d_1a_1+\cdots+d_ma_m+d_{q+1}a_{q+1}+\cdots+d_{q+n}a_{q+n}$($d_1,\ \cdots,\ d_m,\ d_{q+1},\ \cdots,\ d_{q+n}$ 是参数)。如果元素 $w\in L_1$ 且 $w\in L_2$(L_1 和 L_2 与定理 2 中定义相同) , 那么 L_1 和 L_2 是流形 L 的子流形。

证明: 因为元素 $w\in L_1$ 且 $w\in L_2$, 所以由定理 1 可知, 它们的模型可分别表示为: $x=w+b'_1a_1+\cdots+b'_qa_q;\ y=w+b'_{q+1}a_{q+1}+\cdots+b'_ra_r$。既然 $a_1,\ \cdots a_m,\ a_{q+1},\ \cdots,\ a_{q+n}$ 为向量 $a_1,\ a_2,\ \cdots,\ a_r$ 的一个极大线性无关组, 那么 $a_1,\ a_2,\ \cdots,\ a_r$ 中除这 p 个向量外的所有向量都与这 p 个向量线性相关。所以向量 $a_{m+1},\ \cdots a_q,\ a_{q+n+1},\ \cdots,\ a_r$ 可以表示如下:

$$a_{m+1}=e_{m+1,1}a_1+\cdots+e_{m+1,m}a_m+e_{m+1,q+1}a_{q+1}+\cdots+e_{m+1,q+n}a_{q+n}$$
$$\vdots$$
$$a_q=e_{q,1}a_1+\cdots+e_{q,m}a_m+e_{q,q+1}a_{q+1}+\cdots+e_{q,q+n}a_{q+n}$$
$$a_{q+n+1}=e_{q+n+1,1}a_1+\cdots+e_{q+n+1,m}a_m+e_{q+n+1,q+1}a_{q+1}+\cdots+e_{q+n+1,q+n}a_{q+n}$$
$$\vdots$$
$$a_r=e_{r,1}a_1+\cdots+e_{r,m}a_m+e_{r,q+1}a_{q+1}+\cdots+e_{r,q+n}a_{q+n}$$

其中 $e_{m+1,1}$, …, $e_{r,q+n}$ 为参数。将上式代入 L_1 和 L_2 的模型中, 化简后 L_1 和 L_2 可表示如下:

$$x=w+(b'_1+b'_{m+1}e_{m+1,1}+\cdots+b'_qe_{q,1})a_1+\cdots+(b'_m+b'_{m+1}e_{m+1,m}+\cdots+b'_qe_{q,m})a_m+$$
$$(b'_{m+1}e_{m+1,q+1}+\cdots+b'_qe_{q,q+1})a_{q+1}+\cdots+(b'_{m+1}e_{m+1,q+n}+\cdots+b'_qe_{q,q+n})a_{q+n}$$
$$y=w+(b'_{q+n+1}e_{q+n+1,1}+\cdots+b'_re_{r,1})a_1+\cdots+(b'_{q+n+1}e_{q+n+1,m}+\cdots+b'_re_{r,m})a_m+$$
$$(b'_{q+1}+b'_{q+n+1}e_{q+n+1,q+1}+\cdots+b'_re_{r,q+1})a_{q+1}+\cdots+(b'_{q+n+1}e_{q+n+1,q+n}+\cdots+b'_re_{r,q+n})a_{q+n}$$

此时, 可以看出, 当参数 $d_1=b'_1+b'_{m+1}e_{m+1,1}+\cdots+b'_qe_{q,1}$, …, $d_m=b'_m+b'_{m+1}e_{m+1,m}+\cdots+b'_qe_{q,m}$, $d_{q+1}=b'_{m+1}e_{m+1,q+1}+\cdots+b'_qe_{q,q+1}$, …, $d_{q+n}=b'_{m+1}e_{m+1,q+n}+\cdots+b'_qe_{q,q+n}$ 时, 流形 L 成为 L_1。

当参数 $d_1=b'_{q+n+1}e_{q+n+1,1}+\cdots+b'_re_{r,1}$, …, $d_m=b'_{q+n+1}e_{q+n+1,m}+\cdots+b'_re_{r,m}$; $d_{q+1}=b'_{q+1}+b'_{q+n+1}e_{q+n+1,q+1}+\cdots+b'_re_{r,q+1}$, …, $d_{q+n}=b'_{q+n+1}e_{q+n+1,q+n}+\cdots+b'_re_{r,q+n}$ 时, 流形 L 成为 L_2。因此, L_1 和 L_2 是流形 L 的子流形。

1.2 聚类模型

设 D 为一个 d 维的点集, a_1, …, a_d 为张成该空间的正交向量, A 是由 a_1, …, a_d 中的 k 个向量构成的一个 $d\times k$ 矩阵, A 是由其余 $d-k$ 个向量构成的一个 $d\times(d-k)$ 矩阵。

定义 2(聚类模型[9]) 设 μ 是 D 中的一个点, λ 是一个 0 均值 k 维随机向量, 其每个分量(随机变量)取值范围在 $[-R/2,+R/2]$ 之间(R 是数据的范围), Ψ 是一个 0 均值小方差的($d-k$)维随机向量。一个线性流形聚类可表示为:

$$x=\mu+A\lambda+A\Psi \tag{1}$$

由(1)式可知, 一个线性流形聚类具有这样的性质: 线性流形聚类中的每个点都在或者很接近于一个 k 维的线性流形, 且该线性流形由转移向量 μ 和 k 个线性无关的向量描述(A 的列向量, 此后文中称这些向量为流形的特征向量)。由于 λ 是一个 0 均值的随机向量, Ψ 是一个 0 均值小方差的随机向量, 所以 $E(x)=E(\mu+A\lambda+A\Psi)=E(\mu)+E(A\lambda)+E(A\Psi)=\mu+0+0=\mu$。由此可知, 线性流形聚类的中心为 μ。此外, D 中一点 y 到线性流形 L 的法向距离定义为: $D_o(y,L)=\|(I-AA^{\mathrm{T}})(y-\mu)\|$; 点 y 到线性流形 L 的切向距离定义为: $D_t(y,M)=|AA^{\mathrm{T}}(y-\mu)\|$。

2 算 法

由定理 2 及其补充可知, 只要数据中所蕴含的线流形能够满足定理 2 或其补充的条件, 高维流形聚类就可以通过线流形聚类来构造。因此, LSAFCLUS 算法主要分为两个过程: 线流形聚类搜索过程和线性流形聚类融合过程。

2.1 线流形聚类的搜索

线流形搜索过程目的是要找出蕴含在数据中的所有的线流形聚类, 这一过程有两层循环(图 1)。

```
[LineClusterSet, D] = LineManifoldClusterSearching(dataset: D, N, ε, Γ, δ)
While size (D) > N
  X= randomly sample a point from D;
  LineManifold(U_ori, A_ori) = FormingOriginalLineManifold(X); % 初始线流形的形成
  U_old = []; A_old = []; U_new = U_ori; A_new = A_ori;
    While ||U_new - U_old|| > ε
    U_old = U_new; A_old = A_new;
    (U_new, A_new) = UpdatingTranslationVectorandEigenvector(U_old, A_old); % 转移向量和特征向量的更新
    End
  LineManifold(U, A) = LineManifold(U_new, A_new);
  Linecluster= {x| D_o(x, LineManifold) < Γ and D_l(x, LineManifold) < δ};
  LineClusterSet= {LineClusterSet}∪{Linecluster};
  D= D- Linecluster;
End
```

图 1 线流形聚类搜索过程的算法

Fig. 1 The algorithm of line manifold cluster searching

循环中期望给定 4 个参数: N, 在从数据集中剔除掉所找到的线流形聚类后剩余点数目的阈值; ε, 一个很小的常量; Γ, 点到线流形法向距离的阈值; δ, 点到线流形切向距离的阈值。第一层循环是关于剩余数据集(在其中对线流形聚类进行搜索) 的大小。当剩余数据小于所给定阈值时, 算法结束。第二层循环是线流形聚类的优化过程。这一过程主要目的是辨别找到的线流形聚类是不是一个合适的线类。辨别的主要依据是: 如果新的线类和旧的线类之间转移向量的差别小于给定常量 ε, 这个新的线类就可以看作是一个合适的线类。此时, 线流形聚类的优化过程结束, 返回第一层循环。如果这一差别大于给定常量 ε, 那么这一线流形聚类继续优化过程。

在线流形聚类搜索过程中的一个关键问题就是初始线流形的形成。可以通过随机取两点来构造一个线流形。空间近邻信息已经被应用于加快聚类算法的收敛速度和提高算法的判别性能等方面[6-7]。算法也将空间近邻信息引入到初始流形的形成中, 以提高算法收敛的速度和计算结果的准确度。首先, 随机抽样出第一个点, 并将该点作为初始的转移向量。然后找到该点的最近邻, 并将其作为第二个点。最后通过这两个点构造特征向量, 进而得到初始的线流形。

线性流形聚类的中心(类内基因的平均表达水平) 是转移向量, 但是如果用类内基因的平均表达水平作为转移向量, 聚类得到基因表达数据的类别中心的表达模式非常模糊, 进而会使得所得到的基因聚类生物意义不明确[8]。因此算法中选用类内的真实基因作为新的转移向量。至于特征向量的更新, 因为特征向量可以通过线流形上的两点来构造, 因此只需要找到线流形上两个合适的点即可。既然转移向量是在线流形上的点, 那么可以将其作为第一个点, 同时选择法向距离最小的点作为第二个点。

2.2 线流形聚类融合过程

线流形聚类的融合过程以线流形聚类搜索过程的结果为基础。高维流形聚类形成过程实质是一个线流形聚类的层次融合过程, 从 c_1 开始, 直到没有线流形聚类在集合中结束。若线流形聚类搜索过程得到 k 个线流形聚类c_1, c_2, …, c_k, 剩余点集为 D。当对线性流形聚类 c_1 进行融合时, 如果 c_1 与聚类集合中其它的某些类, 不妨设为 c_q 和c_r ($2 \leqslant q < r \leqslant k$), 满足 1.1 节中定理描述的条件, 则 c_1、c_q 和c_r 融合为一个新的高维线性流形聚类c_{k+1}、c_1、c_q 和c_r 从流形聚类集中剔除不再被考虑, 并从剩余点集 D 中剔除那些属于线性流形聚类c_{k+1}的点。当 $2 < q$ 时, 线性流形聚类 c_2 融合时需要考虑的流形聚类集为 c_3, …, c_{q-1}, c_{q+1}, …, c_{r-1}, c_{r+1}, …, c_k, c_{k+1}。当 $q = 2$ 时, 因为线性流形聚类 c_2 已经从流形聚类集中被剔除, 所以不被考虑, 直接考虑下一流形聚类的融合过程。如果 c_1 与类集中所有类都不满足 1.1 节中定理描述的条件, 这说明 c_1 不是其它任何高维流形的子流形, 其本征维数为 1, 是一个线流形聚类,

将其从流形聚类集中剔除后,进行流形聚类集中下一个流形聚类 c_2 的融合过程。此时,c_2 融合时需要考虑的流形聚类集为 c_3, …, c_k。其它流形聚类的融合过程与流形聚类 c_1 的融合过程类似。

3 实验及结果分析

3.1 实验数据

两组基因表达数据用于本文的研究。第一组数据是一个酵母的基因表达数据[9],该基因表达数据是一个时间序列数据,给出了部分基因的相关功能注释。根据它们的功能注释,从中选出6类共335个基因作为基本研究对象,并从剩下的基因中随机抽取部分基因作为噪声基因构成数据集进行实验。在改变噪声比(定义为噪声基因数与已知功能基因数335的比值)的情况下构建了多个数据集进行多次实验,其中,噪声比的变化范围是[0,0.18]。显然,第一组实验数据集可以看作是已知分类的数据集。第二组数据是人类HeLa细胞周期数据[10],包含1095个周期表达的基因在实验条件Thy-Thy3下47个时间点的采样和在实验条件Thy-Noc下19个时间点的采样。将这1095个基因的表达数据作为聚类分析的第二组实验数据,并在对这1095个基因聚类前,先对其进行预处理和标准化。

3.2 算法的参数设置

LSAFCLUS期望4个输入参数:N、ε、Γ 和 δ。在已知分类的基因表达数据的聚类中,N 值取为噪声数。在未知分类的基因表达数据的聚类中,噪声数目未知,N 值取为数据集大小的十分之一。ε 是一个小常量,它对算法的影响很小。当 ε 的值在[10^{-10}, 10^{-7}]之间变化时,多次实验的结果都没有太大的差别,因此设置 ε 值为 10^{-7}。至于 Γ 和 δ 的设置,算法中采用最小误差阈值方法[11]。

3.3 实验结果及分析

3.3.1 已知分类的实验数据

对于已知分类的数据,聚类纯度(CP)被用来度量聚类的准确度。聚类纯度的定义是基于一个混合矩阵。混合矩阵反映了输出的聚类与输入类间的匹配程度。矩阵(i, j)位置上的元素表示同时属于输出聚类 i 和输入类j 的数据点的个数。令 K 表示输出聚类的数目,M 表示输入类的数目,$|D|$ 表示数据集的大小,$|C_i|$ 表示输出聚类的大小,$|C_{ij}|$ 表示同时属于输出聚类 i 和输入类j 的数据点的个数,则 C_i 类的聚类纯度可表示为 $Purity(C_i)=\frac{1}{|C_i|}\max_j(|C_{ij}|)$。因此聚类纯度可表示如下:

$$CP=\sum_{i=1}^{K}\frac{|C_i|}{|D|}Purity(C_i)=\frac{1}{|D|}\sum_{i=1}^{K}\max(|C_{ij}|) \tag{2}$$

CP描述了输出聚类和输入类间的匹配准确度,其值越大,聚类效果越好。图2给出了不同聚类算法在噪声比变化情况下对已知分类的实验数据集的聚类准确度曲线。由图2可以看出,无论噪声比如何变化,对该组实验数据,LSAFCLUS在四种聚类算法中都得到最好的聚类性能,LMCLUS算法的聚类性能较好,K-means次之,而Dbscan的聚类性能最差。显然,两种线性流形的方法比两种传统聚类方法的聚类效果要好。一方面是因为K-means和Dbscan在全维空间中度量对象之间的相似性,而由于高维数据的稀疏性,这种度量方式会失效。另一方面是因为LMCLUS和LSAFCLUS作为线性流形聚类的算法,作为子空间聚类算法的一种,可以发现蕴含在高维数据中的任意方向的数据模块。

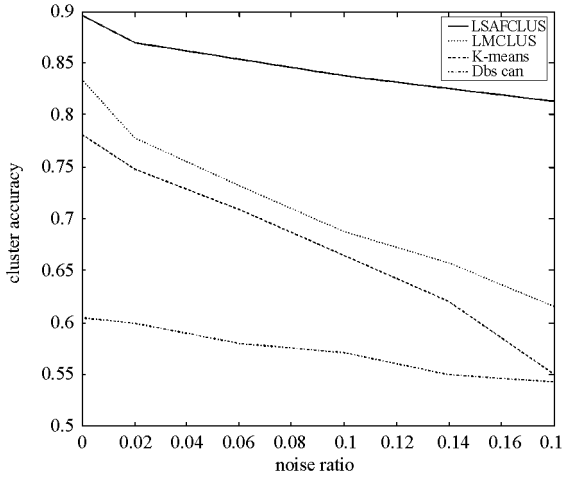

图2 噪声比变化下不同算法的聚类准确度

Fig. 2 Cluster accuracy of different clustering methods under various noise ratios

此外,由图2可以看出,当数据中存在噪声时,

K-means 和 LMCLUS 方法的性能急剧下降, 而 LSAFCLUS 算法仍然能保持很好的聚类准确度。这说明 K-means 和 LMCLUS 这两种方法很容易受噪声数据的影响, 而 LSAFCLUS 方法有较好的抗噪性。这是因为 LSAFCLUS 算法采用点到流形的法向距离和切向距离两者一起作为线性流形的距离度量, 并且充分利用了空间最近邻信息改进了初始流形的形成方法, 很好地克服了噪声的影响。所以 LSAFCLUS 方法适用于对带有噪声的基因表达数据集进行聚类。

3.3.2 HeLa 细胞周期数据

为了进一步验证基于线流形的基因表达数据聚类方法的有效性, 使用该方法分别在 HeLa 数据(Data1)的两个不同的实验条件 Thy-Thy3 和 Thy-Noc 下得到的周期数据上进行聚类分析, 分别得到 10 个和 11 个类别。将聚类结果提交到 Generic GOTermFinder[12] 上得到 p-value 小于 10^{-2} 的类别, 两组实验结果中功能一致的有 8 类。表 1 列出了这 8 类在实验 Thy-Thy3 中所对应的细胞周期阶段(Phase)和主要的 GO 分类。其中, Cluster 表示类别编号, (p-value) 表示该类注释的显著性值。

表 1 8 个聚类在实验 Thy-Thy3 中所对应的细胞周期阶段和 GO 分类

Tab.1 Corresponding cell-cycle phases and GO terms of 8 clusters in Thy-Thy3

Cluster	Phase	Biological Process (p-value)	Cell Component(p-value)
C1	G2/M M/G1	cellular process (2.17E-14) macromolecule metabolic process (2.28E-11) biological regulation (5.04E-06) primary metabolic process (3.90E-09)	nucleus (4.92E-14) nuclear part (1.83E-13) Intracellular (2.66E-11)
C2	G1/S S	DNA metabolic process (7.23E-30) macromolecule metabolic process (4.84E-07)	nucleus (1.25E-12) intracellular organelle (1.48E-11) membrane-bounded organelle (1.05E-09)
C3	S	DNA metabolic process (1.68e-19) DNA repair (1.63E-11) response to DNA damage stimulus (5.75E-19)	chromosome (3.50E-23) nucleus (2.05E-4)
C4	G1/S	DNA metabolic process (3.85E-12) DNA replication initiation (5.34E-09)	nucleus (2.43E-8)
C5	G2 G2/M	Mitosis (3.40e-10) cell division (9.51e-9) nuclear division (6.70E-05)	intracellular non-membrane-bounded organelle (8.26E-9)
C6	G1/S	DNA packaging (1.66E-08) nucleosome assembly (1.18E-06) chromosome organization (2.47E-05)	nucleosome (6.30E-08) protein-DNA complex (5.39E-07)
C7	G2/M M/G1	regulation of kinase activity (5.2E-07) regulation of cellular process (7.36-6)	nucleus (6.41E-9)
C8	M/G1 G1/S	response to stimulus (4.21E-4) response to stress (6.62E-4)	Chromosome (3.50E-13) chromosome, centromeric region (3.36E-12)

由表 1 知, 基于线流形的聚类方法能够得到一些具有显著生物学意义的聚类。C1 和 C2 都参与了细胞核中的大分子代谢的过程。C1 主要发生在 G2/M 和 M/G1 期, 主要参与了细胞骨架的形成, 核膜瓦解, 核仁消失, Golgi 体、ER 等细胞器解体形成小的膜泡, 染色体分离等过程中的大分子代谢。C2 主要发生在 G1/S 和 S 期, 主要参与了 DNA 代谢及其涉及的蛋白质、糖、脂和能量等营养物质的代谢过程。同样, C3 和 C4 都参与了细胞核中的 DNA 代谢, 主要包括 DNA 复制和 DNA 修复。C3 主要发生在 S 期, C4 主要发生在 G1/S 期, 由于在时间上相差不远, 功能上有一定的相似性, 但是调控机制的不同使它们在功能上还是有一定差别的: C3 的重心在于 DNA 的复制及复制前的 DNA 修复, C4 的重心被定位于 DNA 复制的开始和 G1/S 转化点的检验。C5 主要发生在 G2 和 G2/M 期, 主要参与了核分裂、有丝分裂等细胞分裂中的一系列生物过程。C6 主要发生 G1/S 期, 主要参与了 DNA 的打包、核小体装配及染色体的组

织。C7主要发生在G2/M和M/G1期,主要参与了激酶活性调控和细胞过程的调控。C8主要发生在M/G1和G1/S期,主要参与对外界刺激和压力的响应等生物过程。

4 结论

本文针对基因表达数据高维、有噪声等特点,提出了一种新的基于线流形的基因表达数据聚类方法LSAFCLUS。它不同于传统的聚类方法,是一种线性流形聚类方法。该方法基于线流形的搜索,充分运用了空间近邻信息,克服了噪声的影响,保证了算法的鲁棒性;采用真实基因的表达水平作为转移向量的更新策略进行优化迭代,进而挖掘基因表达数据中蕴含的基因模块,提高了算法聚类的准确性。实验结果说明与其它聚类算法相比,该算法在无噪声和带有噪声的基因表达数据集中都能够得到高的聚类准确性,说明它对基因表达数据聚类的适用性。此外,通过对Hela基因表达数据的聚类,该算法可以得到具有显著生物学意义的功能聚类,说明了该算法对于基因表达数据聚类的有效性。当然,我们的研究也还存在其局限性,LSAFCLUS算法是基于线性流形的一种算法,它不能找到蕴含在基因表达数据中的非线性流形。因此,研究能够找到数据中蕴含的非线性流形聚类算法是未来的一个研究方向。

参考文献:

[1] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data[J]. Data Mining and Knowledge Discovery, 2005, 11(1): 5-33.

[2] Alfarano C, Andrade C E, Anthony K, et. al. The Biomolecular Interaction Network Database and Related Tools: 2005 Update[J]. Nucleic Acids Res., 2005, 33: D418-424.

[3] Zanzoni A, Montecchi-palazzi L, Quondam M, et. al. MINT: A Molecular INTeraction Database[J]. FEBS Lett., 2002, 513: 135-140.

[4] Harpaz R, Haralick R. Exploiting the Geometry of Gene Expression Patterns for Unsupervised Learning[C]// Proceedings of the 18th International Conference on Pattern Recognition, 2006: 670-674.

[5] Haralick R, Harpaz R. Linear Manifold Clustering in High Dimensional Spaces by Stochastic Search[J]. Pattern Recognition. 2007, 40: 72-84.

[6] Deng H, Wu Y H, Duan J A. Adaptive Learning with Guaranteed Stability for Discrete-time Recurrent Neural Networks [J]. Journal of Central South University of Technology, 2007, 14(3): 685-690.

[7] Zhou X C, Shen Q T, Liu L M. New Two-dimensional Fuzzy C-means Clustering Algorithm for Image Segmentation[J]. Journal of Central South University of Technology, 2008, 15: 882-887.

[8] 王广云. 肿瘤基因芯片表达数据分析相关问题研究[D]. 长沙: 国防科技大学,2009.

[9] Shapira M, Segal E, Botstein D. Disruption of Yeast Forkhead-associated Cell Cycle Transcription by Oxidative Stress[J]. Mol. Biol. Cell, 2004, 15: 5659-5669.

[10] Whitfield M L, Sherlock G, Saldanha A J, et al. Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors[J]. Molecular Biology of the Cell, 2002, 13: 1977-2000.

[11] Kittler J, Illingworth J. Minimum Error Thresholding [J]. Pattern Recognition, 1986, 19: 41-47.

[12] Generic GOTermFinder[DB]. http://go.princeton.edu/cgi-bin/GOTermFinder.