基于遗传算法的基因分类

蔡立军[1,2], 林亚平[2,1], 卢新国[1], 易叶青[1], 李小龙[1]
(1. 湖南大学计算机与通信学院, 湖南长沙 410082; 2. 湖南大学软件学院, 湖南长沙 410082)

摘 要: 独立分量分析(ICA)是应用于基因分类的一种统计方法. 但独立分量分析中的估计分离矩阵算法主要采用了随机梯度算法、自然梯度算法, 这些基于梯度下降的寻优算法很容易陷入局部极值, 所得结果不精确. 本文提出了一种基于遗传算法的基因分类算法, 其基本思想是利用遗传算法代替独立分量分析中的传统的估计分离矩阵算法, 对基因表达式数据进行分类, 从而克服了结果不精确的问题. 分析和实验结果表明, 该分类方法获得了更好的分类效果.

关键词: 基因分类; 基因表达阵; 独立分量分析; 遗传算法
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2006) 11-2115-05

Gene Clustering Based on Genetic Algorithm

CAI Li-jun[1,2], LIN Ya-ping[2,1], LU Xin-guo[1], YI Ye-qing[1], LI Xiao-long[1]
(1. *School of Computer and Communication, Hunan University, Changsha, Hunan 410082, China;*
2. *School of Software, Hunan University, Changsha, Hunan 410082, China*)

Abstract: Independent component analysis (ICA) is a statistical method applied to gene cluttering. Estimative separation matrix algorithm of ICA uses mainly Random Gradient Algorithm or Natural Gradient Algorithm. And yet these algorithms can only get the partial optimized solution. This paper proposes a new algorithm of gene clustering based on genetic algorithm. The key idea is by using genetic algorithm instead of previous estimative separation matrix algorithms in ICA to classify gene expression data. The former has an advantage of overcoming partial optimized solution. The analysis and experiments support our conclusion that gene clustering based on genetic algorithm has better performance.

Key words: gene clustering; gene expression array; independent component analysis(ICA); genetic algorithm

1 引言

目前, 聚类技术广泛应用于基因表达式数据分析, 通过基因聚类可以将海量的基因表达数据划分成数量相对较少且具有生物意义的组已经成为生物信息学的研究热点.

Raychaudhuri 等在文[1]中利用主分量分析(Principal Components Analysis, PCA)对酵母孢子发芽的数据集[2]进行实验, 发现只要 2 个隐含变量就可以获取该数据集在孢子发芽过程中 7 个时间点所收集的数据绝大部分信息. 最近, Michae[3]等利用奇异值分解(Singular Value Decomposition, SVD)对基因表达数据进行分析, 提出当利用协方差矩阵计算 PCA 时, 利用 SVD 分析和 PCA 分析具有等同的作用. Liebermeister[4]提出了基于独立分量分析(Independent Components Analysis, ICA)[5~7]的基因表达的线性模型, 认为主要的模式与不同的生物功能具有相关性, 对这些模型产生线性影响的系数分量是基于最小统计独立的. Hori 等[8,9]则利用 ICA 方法对基因数据进行盲分类, 并与基于 PCA 的分类方法进行比较, 指出基于 ICA 的盲基因分类取得更好的效果, 更具有可行性. 但独立分量分析中的估计分离矩阵算法主要采用了随机梯度算法、自然梯度算法和联合对角化操作, 这些基于梯度下降的寻优算法很容易陷入局部极值, 所得结果不精确.

本文提出了一种基于遗传算法的基因分类算法, 其基本思想是利用遗传算法代替独立分量分析中的传统的估计分离矩阵算法, 从而克服了结果不精确的问题, 理论分析和实验结果表明该算法是可行的. 实验表明, 采用了该算法处理后获得的基因表达式数据, 与没有处理过的源基因表达式数据相比, 均使用了传统的 K-MEANS 算法和 SOM 算法进行分类后, 前者获得了更好的分类效果; 并且, 与传统的 ICA(包括 FAST-ICA)获得的分离矩阵 $\boldsymbol{W}$ 相比, 基于该遗传算法获得的分离矩阵 $\boldsymbol{W}$ 在用于基因分类时, 能够获得更好的结果.

2 基于 ICA 的基因分类

2.1 ICA 概述

ICA 是近些年发展起来的一种新的统计方法[7]. 该方法

收稿日期: 2006-05-13; 修回日期: 2006-08-06
基金项目: 湖南省自然科学基金(No. 06JJ20049)

是将观察到的数据进行某种线性分解, 使其分解产生相互统计独立的分量. ICA 模型可以描述为:

$$\boldsymbol{X}=\boldsymbol{AS} \tag{1}$$

$\boldsymbol{X}$, $\boldsymbol{S}$ 的列向量分别是观测信号和源信号, $\boldsymbol{A}$ 为混合矩阵. 该模型揭示了被观察到的信号如何由源信号混合而成. ICA 假设源分量之间是统计独立的, 并且这些分量属于非高斯分布. 混合矩阵 $\boldsymbol{A}$ 为方阵, 并有逆矩阵 $\boldsymbol{W}$, ICA 模型可以用下式表示:

$$\boldsymbol{Y}=\boldsymbol{WX} \tag{2}$$

ICA 方法是一种盲源信号分离(Blind source separation, BSS)方法. 这里"source"指的是原始信号即独立分量, 像"鸡尾酒会"问题中的说话者;"blind"指对混合阵几乎未知, 对原始信号进行很少的假设. 给定 M 个混合信号, ICA 能同时估计出 M 个分量或 $K\leq M$ 个分量, ICA 是 BSS 中使用最广泛的方法之一.

2.2 基于 ICA 的基因分类算法

基因表达式数据是通过在不同观测时间点或试验条件下获得的细胞样本数据, 通常基因表达式数据用矩阵形式保存, 令 X 表示一个 $m\times n$ 的基因表达阵, 矩阵第 i 行对应于第 i 个基因, 第 j 列对应于第 j 个观测样本, 而矩阵的每个元素 x_{ij} 纪录了第 i 个基因在第 j 个观测样本中的表达水平, 通常 m 要远大于 n.

基于 ICA 的基因分类的基本思想分为两步, 首先利用独立分量分析(ICA)对基因表达式数据进行预处理, 去掉基因表达式数据之间的线性关系, 然后设计分类器对基因表达阵的基因数据进行分类. 但独立分量分析中的估计分离矩阵算法主要采用了随机梯度算法、自然梯度算法和联合对角化操作, 这些基于梯度下降的寻优算法很容易陷入局部极值, 所得结果不精确.

3 基于遗传算法的基因分类

3.1 遗传算法的基因分类可行性分析

基因的表达水平是通过隐含变量的线性组合来决定的, 并且基因表达式的这种线性模型与不同的生物功能具有相关性, 对这些模型产生线性影响的系数分量是基于最小统计独立的. 基于 ICA 的基因分类算法, 本质上是将基因表达矩阵的列看成已混合信号, 然后进行盲源信号分离, 对于盲源信号分离的问题, 可以简单地认为在观测信号 X 的基础上寻找一个分离矩阵 $\boldsymbol{w}=[w_1, w_2, \cdots, w_n]^{\mathrm{T}}$, 并计算出 $\boldsymbol{Y}=\boldsymbol{WX}$ 使得 $\sum_{i=1}^{n}|cum(y_i:4)|$ 取得极大值(对于四阶累计量 $cum(y_i:4)$ 取绝对值的原因是考虑到源信号中亚高斯信号和超高斯信号同时存在, 对于源信号中含有高斯信号 $cum(y_i:4)$ 的值为 0), 为了讨论的方便, 不妨令代价函数:

$$f(w)=\sum_{i=1}^{n}|cum(y_i:4)|=\sum_{i=1}^{n}|cum(w_ix_i:4)| \tag{3}$$

对于上述代价函数可以得出如下的结论:

结论 1 当代价函数取得极大值时, 当且仅当 $\boldsymbol{Y}$ 满足 $cum(y_i:2)=1$ 的约束条件.

结论 2 当代价函数取得极大值时, 对于每一个输出信号 y_i 有且仅有一个 w_i 使得 $y_i=w_ix$ 成立.

首先证明结论 1.

证明:

假设存在一个 i 使得 $cum(y_i:2)\neq 1$, 不妨令 $cum(y_i:4)$ 取得极值时 $w_i=\xi$, 于是任取 σ 对 ξ 进行扰动.

令 $\tilde{\xi}^4=\xi^4+\sigma(cum(y_i:2)-1)^2$ (4)

由累计量的性质可得:

$$\begin{aligned}
f'(\xi)&=cum(y_i:4)\\
&=E(y_i^4)-3cum(y_i:2)^2\\
&=\xi^4E[x^4]-3\xi^4(cum)(x:2)^2\\
f'(\tilde{\xi})&=cum(\tilde{y}_i:4)\\
&=E(\tilde{y}_i^4)-3cum(\tilde{y}_i:2)^2\\
&=(\xi^4+\sigma(cum(y_i:2)-1)^2)E[x^4]\\
&-3(\xi^4+\sigma(cum(y_i:2)-1)^2)cum(x:2)^2\\
f'(\tilde{\xi})-f'(\xi)&=\sigma(cum(y_i:2)-1)^2E[x^4]\\
&-3\sigma(cum(y_i:2)-1)^2cum(x:2)^2\\
&=\sigma(E[x^4]-3cum(x:2)^2)(cum(x:2)-1)^2\\
&=\sigma cum(x:4)(cum(y_i:2)-1)^2
\end{aligned} \tag{5}$$

式(5)中的 $cum(x:4)$ 由观测信号确定(只要 x 不为高斯信号则有 $cum(x:4)\neq 0$), 由 σ 的任意性可知, 必定可取得一 σ 使得式(5)大于 0, 由于代价函数为一累加函数, 所以在不满足约束条件 $E\{\boldsymbol{yy}^T\}=\boldsymbol{I}$ 下取得极大值是不可能的, 所以结论 1 成立. [证毕]

下面给出结论 2 的证明.

证明:

既然有结论 1 作保证那么当代价函数取得极大值时至少有一个 w_i 使得 $y_i=w_ix$. 下面将证明存在两个不同的 w_i 使得代价函数取得极大值是不可能的.

假设在代价函数取得极大值时, 对于某一输出信号 y_i 存在两个不同的 w_i 使得 $y_i=w_ix$ 成立, 不妨令 $\xi_i\neq\xi_j$ 使得 $cum(y_i:4)$ 取极值, 即 $cum(\xi_ix:4)=cum(\xi_jx:4)$ 成立.

由于代价函数为一累加函数, 所以代价函数取得极大值时, 对于每一个累加项应取得极值, 由结论 1 可知 $cum(y_i:2)=1$, 所以

$$\frac{cum(\xi_ix:4)}{cum(\xi_ix:2)^{3/2}}=\frac{cum(\xi_jx:4)}{cum(\xi_jx:2)^{3/2}} \tag{6}$$

由四阶累计量的性质上式变为:

$$\frac{\xi_i^4cum(x:4)}{\xi_i^3cum(x:2)^{3/2}}=\frac{\xi_j^4cum(x:4)}{\xi_j^3cum(x:2)^{3/2}} \tag{7}$$

即有:

$$\xi_i=\xi_j \text{ 成立.}$$

这与假设矛盾, 所以结论 2 成立. [证毕]

3.2 基于遗传算法的盲基因分类算法

3.2.1 参数编码及初始群体的设定

对分离矩阵 W 进行参数编码, 例如对 n 路源信号进行分离, 则分离矩阵为 $n\times n$ 的方阵, 构成的染色体为 $22bit\times n\times n$, 编码如下所示:

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{14} \\ w_{21} & w_{22} & \cdots & w_{24} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

$$= (w_{11}, w_{12}, \cdots, w_{nn})$$

通过随机方式产生由若干个分离矩阵构成的初始群体. 考虑到运算量较大, 初始群体的规模设定为 50 个随机矩阵.

3.2.2 适应度函数的选取　*K-L* 散度、互信息和负熵等多种参数都可以作为信号非高斯性的判据. 更具基因数据的特性. 这里选用四阶统计量作为信号非高斯性的判据, 其适应度函数定义为:

$$J(y) = \sum_{i}^{n} |cum(y_i: 4)| \qquad (8)$$

在 $E\{yy^T\} = I$ 约束条件下, 对于某一分离矩阵 w, $J(y)$ 越大表明, y_i 的独立性越强, 表明分离效果越好.

3.2.3 选择、交叉和变异操作　选择操作, 计算每个个体的选择概率 $P_i = f_i / F$ 及累计概率 $q_i = \sum_{j=1}^{i} p_j$, 选择方法采用轮盘赌, 旋转 m 次即可选出 m 个个体来. 在计算机上实现的步骤为: 产生 $[0, 1]$ 的随机数 r, 若 $r < q_1$, 则第一个个体入选, 否则第 i 个体入选, 且 $q_i - 1 < r < q_i$.

交叉操作, 首先对每个个体产生 $[0, 1]$ 间的随机数 r, 若 $r < p_c$ (p_c 为选定的交叉概率), 则该个体参加交叉操作, 如此选出交叉操作的一组后, 随机配对; 然后对每一配对, 产生 [1, length(染色体)] 间的随机数以确定交叉的位置.

变异操作, 对每一串中的每一位产生 $[0, 1]$ 间的随机数 r, 若 $r < p_m$ (p_m 为变异概率), 则该位变异; 实现变异操作, 即将原串中的 0 变为 1, 1 变为 0.

3.2.4 约束条件的保持　由四阶累计量的定义和性质可知, 四阶累计量作为非高斯性度量的前提条件是零均值, 由结论 1 可知四阶累计量要想取得极值, 就必须满足 $cum(y_i: 2) = 1$ 的约束条件, 因此在计算随机变量的四阶累计量时就必须引入两个基本的操作: 中心化操作和白化操作.

中心化的目的是为了让随机变量满足零均值, 实现起来较为简单, 只需将随机变量减去其均值即可.

白化操作的目的就是要让随机变量 x 通过线性变换 $\boldsymbol{Y} = \boldsymbol{QX}$ 使得随机变量满足 $E\{\boldsymbol{yy}^T\} = \boldsymbol{I}$ 这一约束条件.

由上述分析可知, 主成分分析 (PCA) 恰好能满足上述过程, 所以选取主成分分析 (PCA) 作为白化操作. 主成分分析 (PCA) 不仅能满足约束条件, 还可以完成随机变量的不相关处理, 同时能起到降维作用, 以保证观测信号的数量和源信号的数量相同.

3.3 算法描述

由上述分析可知, 基于遗传算法的盲基因分类可描述为:

Step1: 对基因表达阵列信号进行白化和中心化操作;

Step2: 随机方式产生由 50 个分离矩阵构成的初始群体;

Step3: 由分离矩阵根据公式 $\boldsymbol{Y} = \boldsymbol{WX}$ 计算出 Y

Step4: 对 Y 进行白化和中心化操作;

Step5: 调用适应度函数计算出每一个染色体的适应度, 并评价个体的适应度;

Step6: 是否达到结束条件, 如果是, 则输出信号转 step9;

Step7: 对染色体进行交叉变异操作产生新一代群体;

Step8: 转 step3;

Step9: 利用 K-MEANS 和 SOM 算法对基因进行分类, 并计算出相关的数据;

Step10: 结束.

4 算法性能分析与仿真实验

定义 1　(多重相关性) 在一个基因表达矩阵中如果样本 (列) 与样本之间存在着相关性则称这种相关性为多重相关性.

性质 1　多重相关性对基因数据矩阵的分类与聚类分析是有害的.

证明: 在聚类和分类分析的过程中主要通过评价各模式点间的相似度和差异, 为了描述的方便, 假定以欧氏距离作为基本测度:

$$d^2(\boldsymbol{e}_i, \boldsymbol{e}_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

其中模式点 $\boldsymbol{e}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \in R^P$, 即 $\boldsymbol{e}_i$ 用 p 个样本点来描述.

现假设所要评估的系统只有两个主要性质 Y_1, Y_2, 但在采样的过程中, 却得到 9 个完全相关的 x_1, x_2, …, x_9 变量来描述 Y_1, 而对另一个性质 Y_2 却仅用一个变量 x_{10} 来表示, 则事实上有:

$$d^2(e_i, e_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

$$= \sum_{j=1}^{9} (x_{ij} - x_{kj})^2 + (x_{i10} - x_{k10})^2$$

显然, 性质 Y_1 的信息在计算中被重复了 9 次, 起到了绝对重要的影响作用, 其它测量度与此类似, 故性质 1 的结论成立. [证毕]

性质 2　利用 ICA 先对基因表达矩阵进行去相关性操作然后再进行聚类分析比直接对基因表达矩阵进行聚类分析效果要好.

由性质 1 保证, 性质 2 显然成立.

本文引入遗传算法求解混合矩阵, 遗传算法比传统的优化算法的优化能力要强, 因此在消除样本间的多重相关性比 ICA 算法要强, 在此基础上的分类效果将会更好.

为了进一步验证基于遗传算法的分类模型和分类效果, 我们采用了由 Chu1 等收集的酵母基因数据集进行实验. 该数据集包含了酵母基因组中的 6118 个基因表达数据, 这些数据是在酵母发芽过程中 0.0、0.5、2.0、5.0、7.0、9.0 和 11.5 小时 7 个时间点获得. Chu 等将该数据集中的酵母基因分为 7 个类. 首先他们定义了 7 个小类, 每个小类中包含有 3-8 个不同时期具有代表意义的基因, 手工提取获得了这 7 个类别的平均表达式 (图 1).

然后在此基础上根据基因与平均模型的相似度将其余的基因分到不同的类中.

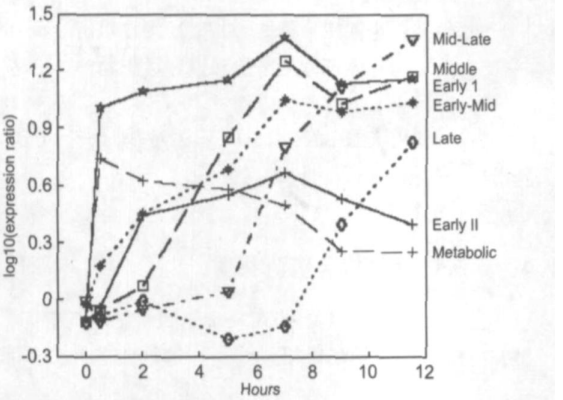

图 1 酵母菌 7 种类别基因的平均表达式

实验 **1** 我们首先采用 K-MEANS 和 SOM 算法对这 6118 个基因表达式数据进行分类. 在 K-MEANS 算法中, 我们预定义 $K=7$, 分类之后计算人工分得的 7 类的平均值如图 2; 在 SOM 算法中, 神经元的数目为 1× 7, 神经元的向量的值如图 3.

说明: 由于基因表达式向量中的数据存在着相关性, 导致了采用 K-MEANS 算法和 SOM 算法分类之后得到的结果与手工分类相差较远.

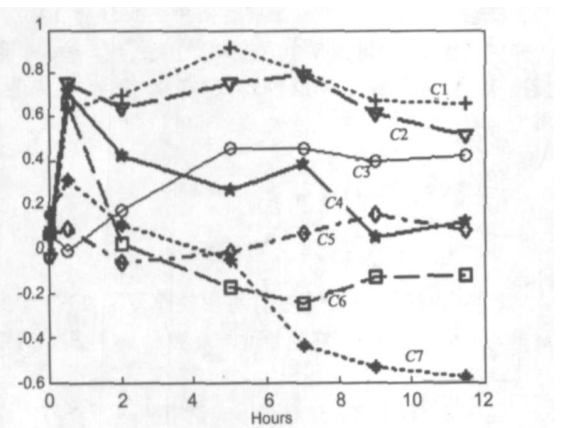

图 2 K-MEANS 分成 7 类时基因的平均表达式

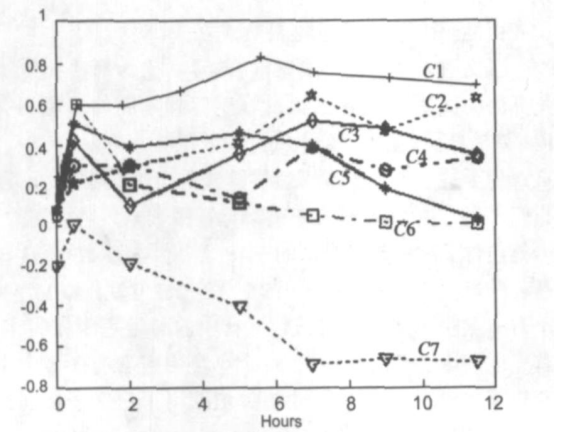

图 3 SOM 分成 7 类时基因的平均表达式

实验 2: 我们接着对基因表达式进行独立化处理. 图 5 是采用了 FastICA 进行独立化处理, 图 4 是使用遗传算法进行独立化处理, 取 $p_c=0.5$, $p_m=0.01$, 将所得结果进行 K-MEANS 分类.

说明: 由于 FastICA 采用梯度算法求解混合矩阵其优化能力较差并不能很好的消除基因表达式数据间的多重相关性, 因此通过 FastICA 处理后再分类的效果比直接用 K-MEANS 的效果没有明显的提高, 由于遗传算法解混合矩阵时更精确, 因此通过遗传算法处理后分类效果更好.

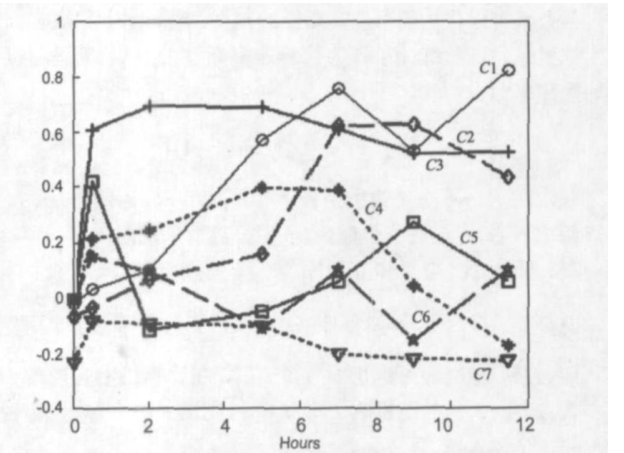

图 4 GA 处理后使用 K-MEANS 分成 7 类时基因的平均表达式

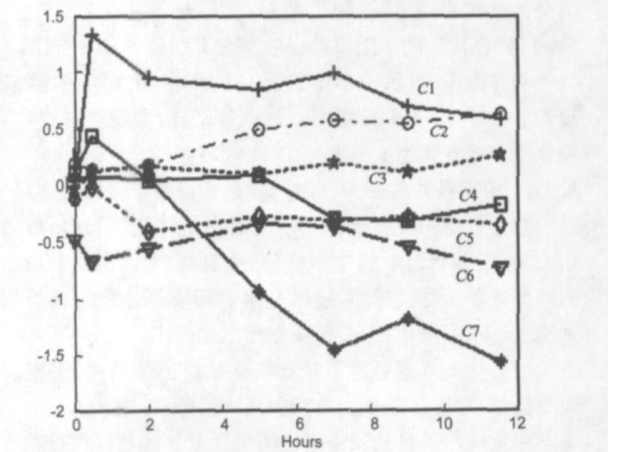

图 5 FastICA 处理后使用 K-MEANS 分成 7 类时基因的平均表达式

5 结论

基因分类问题是一个公开的富有挑战性的问题, 本文提出了一种基于遗传算法的基因分类算法, 其基本思想是利用遗传算法代替独立分量分析中的传统的估计分离矩阵算法, 从而克服了结果不精确的问题, 它具有比传统 ICA 更强的去多重相关性能力, 理论分析和实验结果表明该算法是可行的. 实验表明, 采用了该算法处理后获得的基因表达式数据, 与没有处理过的原基因表达式数据相比, 均使用了传统的 K-MEANS 算法和 SOM 算法进行分类后, 前者获得了更好的分类效果; 并且, 基于该遗传算法与传统的 ICA(包括 FAST ICA) 获

得的分离矩阵 W 相比, 前者获得的分离矩阵 W 在用于基因分类时, 能够获得更好的结果.

参考文献:

[1] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series[A]. Pacific Symposium on Biocomputing[C]. Honolulu, Hawaii, USA, 2000. 452- 463.

[2] Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown P O, I Herskowitz. The transcriptional program of sporulation in budding yeast[J]. Science, 1998, 282(3): 699-705.

[3] Michael E W, Andreas R, Luis M R. Singular value decomposition and principal component analysis[EB/OL]. http://public.lanl.gov/mewall/kluwer2002.html, 2005-03-08.

[4] Liebermeister W. Linear modes of gene expression determined by independent component analysis[J]. Bioinformatics, 2002, 18(1): 51- 60.

[5] Hyvärinen A. Survey on independent component analysis[J]. Neural Computing Surveys, 1999, 23(2): 94- 128.

[6] Hyvarinen A. Fast and robust fixed-point algorithm s for independent component analysis[J]. IEEE Trans. on Neural Networks, 1999, 10(3): 626- 634.

[7] 杨竹青, 李勇, 胡德文. 独立成分分析方法综述[J]. 自动化学报, 2002, 28(5): 762- 773.

[8] Hori G, Inoue M, Nishimura S, Nakahara H. Blind gene classification: an application of a signal separation method[DB/CD]. Proc Genome Informatics Workshop (GIW2001), Tokyo, Japan, 2001. 255- 256.

[9] Hori G, Inoue M, Nishimura S, Nakahara H. Blind Gene Classification: An ICA-based Gene Classification/Clustering Method-RIKEN BSI BSIS Technical Report No. 02-5, 2002 [EB/OL]. http://www.bsp.brain.riken.go.jp/~hori/BSISTR/BSISTR-02-5.pdf, 2005-08-16.

作者简介:

蔡立军　男, 1964 年出生于湖南常德, 湖南大学计算机与通信学院博士研究生, 副教授, 主要研究方向: 机器学习、基因分类.

E-mail: ljcai@hnu.cn

林亚平　男, 1955 年出生于湖南邵阳, 湖南大学教授, 博士生导师, 主要研究方向为计算机网络, 机器学习. E-mail: yplin@hun.cn