一种新型多类别生物芯片 cDNA 基因表达数据标准化方法

吕建平[①] Wang Yue[②]
[①](苏州大学电子信息学院 苏州 215021)
[②](*Virginia Polytechnic Institute and State University*, *Arlington VA*22203, *USA*)

摘 要：cDNA 生物芯片表达数据广泛用于生物医学研究，利用计算机对其进行处理还有很多挑战性课题。该文提出了一种新的基于不变基因的多类生物芯片监督型集合 cDNA 表达数据标准化方法。在达到标准化的同时，该方法也可直接用于基因表达数据的特征选择，实验证明效果较好。

关键词：DNA 基因；跨类；标准化；多类多样本；特征选择
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2009)06-1350-04

A Cross-phenotype Normalization Method for cDNA Gene Expression Data

Lü Jian-ping[①] Wang Yue[②]
[①](*School of EE & Information*, *Soochow University*, *Suzhou* 215021, *China*)
[②](*Virginia Polytechnic Institute and State University*, *Arlington VA*22203, *USA*)

Abstract: cDNA microarray expression data is widely used to help biomedical research. There are so many challenges when the computer and pattern recognition methods are used to process and analysis the data. In this paper, a novel hybrid cross-phenotype normalization method is proposed which deals with supervised multi-class multi-sample cDNA expression data set based on invariantly expressed genes. The algorithm can be directly used as a feature selection method for gene classifier. The result is satisfactory.

Key words: DNA gene; Cross-phenotype; Normalization; Multi-class and multi samples; Feature selection

1 引言

cDNA 生物芯片数据广泛用于生命科学研究。每个组织样本所包含的生物芯片基因数可达数万，哪些基因与疾病有关，其相互之间的关系如何，如何利用计算机来处理是研究人员最关心的问题。

利用模式识别处理数据，首先要对需分类数据进行标准化处理。数据的标准化就是将各个数据按比例缩放或通过函数变换将其映射到某个小的数值区间[1]，以便可比较可计算。

生物研究人员在获得各样本基因表达数据的过程中，由于存在具体实验操作环境不一致而导致基因表达量(intensity)的变化，生命组织和疾病类型差异以及所用生物芯片的差异，可使各个样本以及平行实验的数据处于不相同的水平。为获取精确的生物信息，滤除表达数据中的实验噪声，消除基因芯片内染色偏差和探针引起的空间差异，改进不同基因芯片间的探针标记误差，浓度误差，杂交效率等各种因素而对此数据处理的过程称为基因表达数据标准化[2]。本文所言的数据标准化指对样本的cDNA基因表达数据的标准化，此研究是一项具备挑战性的工作。对此，当前学术界已经提出很多方法，例如，线性回归(Linear Regression, LR)，位序不变(Invariant Ranking, IR)，分位数(Quantile)，非线性迭代回归(Iterative Nonlinear Regression, INR)等方法，还有些研究人员提出利用小波(Wavelets)，支持向量机(SVM)等方法完成非线性回归运算。以上各算法可参见文献[2, 3]。

对于上面提到的各种标准化处理方法，一般含有两个步骤：(1)选定参与系数计算的基因，(2)选择具体的线性或非线性回归方法。除个别算法(例如分位数方法)，大部分标准化方法局限于将单个样本X回归到单个样本Y的范围，难以处理多类-多样本型集合数据的情况。

本文讨论了一种基于不变基因(数据表达不变的基因，Invariantly Expressed Genes, IEGs)的监督型多类(生物类型≥2)多样本(每类样本数≥2)基因表达数据的组合型标准化方法。然后根据找到的不变基因集合，完成上述数据集合标准化工作。

本文分为以下各部分：(1)讨论了不变基因和监督型多类多样本数据集合的概念；(2)讨论了如何利用分位数方法以获得类内的虚拟代表样本；(3)讨论了多维散点图以及如何利用该种散点图和 INR[4]方法找到不变基因集合的算法步骤，完成跨类型基因数据集合标准化(Cross-Phenotype Normalization, CPN)；(4)给出实验结果以及结论，说明本

2008-10-24收到，2009-01-12改回
美国 NIH 的 CA109872 项目基金资助课题

算法与特征选择的关系。

2　监督型多类多样本数据集合标准化和不变基因

2.1 多类多样本监督型数据集合

若所获数据为非监督型，处理起来困难较大。本文讨论的是监督型数据，即已知总的生物种类，以及各样本集合的具体归属类。

2.2 不变基因与非不变基因

生物学已有研究结果。对于各种生物而言，其大部分的基因相同或相似。例如，国外上百位科学家已在英国《自然》杂志上联合宣布，他们成功破译了老鼠基因组序列。研究发现，老鼠体内约有 3 万个基因，数量与人类基因接近，其中绝大部分相同，人类与老鼠共享着 80%的遗传物质和 99%的基因，了解老鼠非常有助于了解人类自身。其基因组草图显示，老鼠的 20 对染色体上共有约 25 亿个碱基对，与人类 23 对染色体上的 29 亿个碱基对相当接近，DNA 链上基因之间的“空白”片断也非常相似。

与正常人之间相比较，某种基因疾病病人也只会有少数基因产生变化。那些数据表达相同的基因称之为不变基因(IEGs)，而产生了变异或导致生物体不同的基因，称之为非不变基因(Non-IEGs)，非不变基因的数据表达应不相同。

2.3 全局型和局部型标准化方法

所谓全局型标准化方法是指在完成回归系数的计算时，使用待处理样本的全部基因表达数据；而局部型标准化方法使用待处理样本的部分基因表达数据完成回归系数的计算。

对于数据标准化而言，最终结果的评估标准非常重要。早期某些算法[5]在两个样本间进行数据标准化时，所使用的一些评判参数如整体均方差等，希望它越小越好，此观点不够全面。例如，对于不同种类所属的样本数据，其本质上就存在着一些不同的基因数据点，若经标准化处理后，两者全局相似程度越接近，表示引起种类差异的那些基因数据点的表达数据也接近，后续处理就越困难，因此，这不是一种好现象。对于这种情况，全局型标准化方法并不适用。常用的全局型标准化方法有Affymetrix 的 LR 方法[5] ，Loess 方法[6]等。这些方法适用于纠正类内各样本间的整体性数据偏差或操作偏差。

对于跨类样本数据，常用局部型标准化方法。主要有INR[4]，IR[7]等，均使用不变基因(IEGs)进行回归系数的计算。

由此，本文基于如下假设：

(1)对于不同生物种类样本，必然存在部分基因数据导致它们产生生物现象差异，对此，局部型方法比较适合，关键是要找到与这些生物种类无关的基因。

(2)对于同种生物种类，除误差和干扰外，类内各样本数据的分布在很大概率上应为一致。对此种数据，全局或局部型标准化方法均可使用。

(3)基于(2)，对于每个多样本生物种类，我们可以找出一个虚拟的统计样本模型，用来作为该类的代表来参与跨类数据标准化。

(4)若样本基因数据分析只局限于同种生物种类，应使用类内标准化方法(Within-Phenotype Normalization, WPN)，此类方法已有较多论文[2,3]讨论。若样本基因数据分析工作为多生物种类类型，需使用基于局部型方法的跨类数据标准化，例如本文的CPN方法。

2.4 高维散点图

在两个样本间进行标准化处理时，常使用二维散点图，其二维平面中的每个点对(a,b)反映了参与处理的样本 A 和样本 B 中同一基因的值。若进行 3 个不同类样本数据的标准化，可用三维散点图，同理，三维立体空间中的每个点对(a,b,c)反映了参与处理的样本 A，样本 B 和样本 C 中同一基因的值，见图 1；多于 3 样本时可类推。

图 1 中箭头为三维散点图的对角线，连接原点(0,0,0)至基因点(a,b,c)的连线与对角线间有一个“β”角，该连线可以投影到立方空间的 3 个二维平面上。各条投影线与其对应二维平面对角线有一个“α”角，例见图 2。

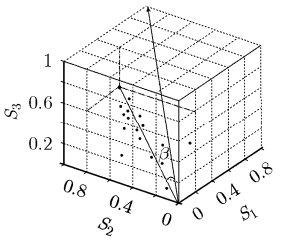

图 1 3D 散点图示意

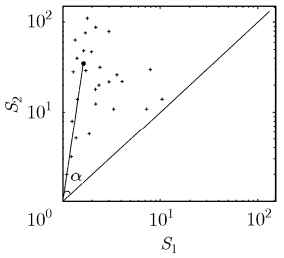

图 2 2D 散点图数据标准化示意

图 2 中黑点是所有基因点(以“+”表示)的均值点，将它与原点(0,0)作一联线，此联线与 2 维对角线也有一个“α”角。从概率角度分析，可见“α”角越大，越需要对有关样本进行标准化处理工作。

注意到：

(1)跨类数据样本中的不变基因 IEG 的分布应围绕着对应散点图空间的对角线。

(2)3D 对角线上的某一点必在其所有各个 2D 投影面对角线上。

(3)3D 空间内某点(a,b,c)趋向其 3D 对角线的行为可以分解为相应各个 2D 投影平面上的移动步。

由此，有第 4 节介绍的组合型跨类数据标准化 CPN 算法。

3　获得某一类的统计代表样本

若某 d 维数据类由 n 个样本组成，这些样本代表不同的测试对象，基于第 2.3 节假设(2)，可以先使用分位数方法[8]以获得各类类内虚拟统计代表样本，然后再使用这些代表样

本参与跨类数据标准化 CPN 方法的运算。所以说，本文讨论的 CPN 方法是一种组合方法。

4 CPN 算法

下面讨论基于 INR 方法[4]的跨类数据标准化方法 CPN。

对于 2 类数据，可命名相应 CPN 算法为 CPN-2；对于 n 类数据，则为 CPN-n。同样，可将 n 类数据所具有的共同不变基因集命名为 IEGs-n。

4.1 基本算法概念

本 CPN 算法是 INR 算法[4]一种扩展。所谓 INR 算法，适用于 2 样本之间的数据标准化，方法是先找出散点图中围绕在 2D 对角线周围较小范围内的基因，作为不变基因，进行一次数据标准化；然后对新产生的数据，作出新的 2D 散点图，再找出此时围绕在 2D 对角线周围较小范围内的基因作为新的不变基因，再进行一次数据标准化；周而复始，直至所选定的不变基因集合趋于驻点，基本不再发生变化，此即为最终的不变基因集合。利用它可以完成最终的数据标准化工作。

本算法中，以三维为例，先找到 3 个投影面中具备最大“α”角的，完成一次 INR 运算。由于处理数据的变化，对于各平面中形成的新的数据分布，重复前面步骤，直至 3 个平面“α”角之和基本不再变化，围绕对角轴的基因部分即为 IEGs-3。

最后再利用常规的线性回归式进行基于 IEGs-3 的数据标准化工作。

4.2 算法步骤

由于 CPN 算法完全基于 INR 算法[4]，而 INR 算法需要使用不变基因集 IEGs，所以在开始 CPN 算法前，需要确定初始的不变基因集 IEGs。本文首先使用文献[2]中叙述的算法，从 $\pi/2$ 开始，使用有限步数，人为地将扇区逐渐缩小靠近对角线，得到扇区角度参数“$\varepsilon 4$”，该参数的含义即为前述的“对角线周围较小范围”(由于 CPN 算法是 INR 算法[4]的扩展，所以沿用了其中的符号“$\varepsilon 4$”)。CPN 运算开始时，可将该范围内的基因作为初始不变基因集 IEGs。

若所获 3 类数据的 3 个 D 维类代表样本表示为 $X_i=\{x_{i1},x_{i2},x_{i3},\cdots,x_{id}\}$，其中 i=1,2,3，为类别号。它们将用作 CPN-3 算法初始时的“当前样本”。

步骤 1 计算当前 3 个代表样本的 X_1-X_2 平面，X_1-X_3 平面，X_2-X_3 平面上各基因点的均值，以及各均值点在该平面对应的各个“α”角。

步骤 2 计算所有各个“α”角之和，若此值(1)与之前各次运算所得的该值相比变化很小，表示趋于它的驻点；或，(2)小于预先确定的某个很小的角度值，则认定围绕 3D 对角线一定范围“$\varepsilon 4$”内的基因可看成 IEGs-3，转向步骤 5；否则，转向步骤 3。

步骤 3 找具备最大“α”角的平面，设为 X_i-X_j 平面。

步骤 4 利用 INR 方法[4]，在 X_i-X_j 平面上找到对应的 IEGs_2，完成一次 X_i-X_j 样本间的 2D 数据标准化。由于形成了新的数据，新的 X_1-X_2 平面、X_1-X_3 平面、X_2-X_3 平面因而形成。即得到新的“当前数据”。转向步骤 1。

步骤 5 根据所得的 IEGs-3，计算回归式 $Y_i=a_{ij}X_{ij}+b_{ij}$ 的系数 a_{ij} 和 b_{ij}，其中 j 是类 i 的样本编号。i=1, 2, 3，为类号。Y_i 是类 i 的已经过上述 CPN 标准化虚拟样本代表。X_{ij} 是类 i 的第 j 个代表样本。最终，利用公式 $X_{ij\text{-new}}=a_{ij}\cdot X_{ij-\text{old}}+b_{ij}$ 完成对每个样本的数据标准化。算法结束。

4.3 算法收敛性和停止

由上可见，CPN 算法是由一系列 INR 算法[4]构成，其实质是完成了高维空间下一系列二维子空间对应基因样本的数据标准化，它是一种扩展的 INR 算法。关于 INR 算法的收敛性，在文献[4]中已有讨论。INR 算法的“$\varepsilon 4$”角度参数的选择将决定其运算结果。CPN 算法基于 INR，“$\varepsilon 4$”角度参数同样重要。对于不同的原始 cDNA 数据集分布，“$\varepsilon 4$”角度参数将不同，需要根据发散情况进行具体调整。INR 发散的现象表现为所得到的 IEGs 集合为空，此时 CPN 也无法计算。

对于本文 CPN 算法而言，只需考虑最终停止条件。在本文中，只要各个 2 维平面的“α”角之和不再减小，算法即停止。

5 结果

作为测试，本文算法使用了美国国家儿童医疗中心提供的cDNA基因样本表达数据。

数据集共有15类121个样本，代表15种不同病型。每样本含22215个基因。这里随机选用了3类样本数较多的数据作为输入：第4类，10个样本；第11类，14个样本；第15类，21个样本。在图3和图4的座标名称上，分别表为样本 C_4，样本 C_{11}，样本 C_{15}。

图 3(a)为所选数据进行 CPN 之前的经过 WPN 处理后的 3 个代表基因 C_4，C_{11}，C_{15} 的初始 3D 散点图及其在 2D 上的投影(图 3 的 3(b)，3(c)，3(d))；图 4(a)为对上述数据进行 CPN 处理之后的 3D 散点图及其在 2D 上的投影(图 4 的 4(b)，4(c)，4(d))。对照其数据分布坐标值，可见经过 CPN 算法处理之后，三类基因的代表数据的分布相对于图 3 而言，图 4 更近地围绕于 3D 散点图对角线，表示各类的代表样本的平行实验表达数据处于相对类似的水平，达到标准化的目的。对于本实验表达数据的 22215 个基因中，最终所获得的这些数据的 IEGs-3 有 22174 个，其他的有 41 个。

本文还试验了该数据集中其他样本类，其 IEGs-3 的数量为 21950 至 22180。所得图形也具有如图 3 和图 4 的性质。由于数据集中存在一个奇异基因点，因而所显示图的各维坐标轴不等长。给出各 2D 投影后，容易理解算法结果。

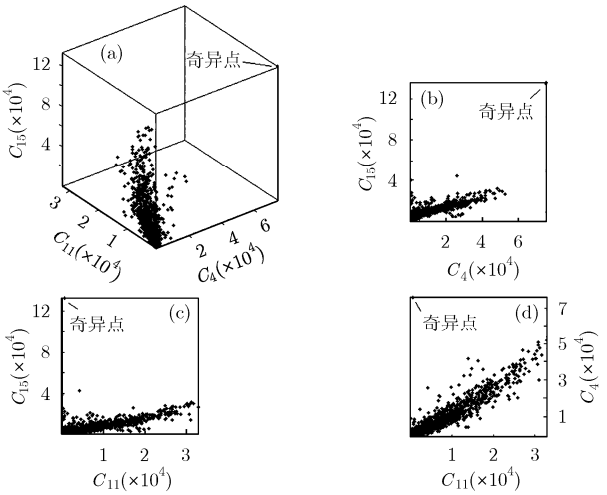

图 3 未经 CPN 处理的 3D 散点图及其各 2D 投影

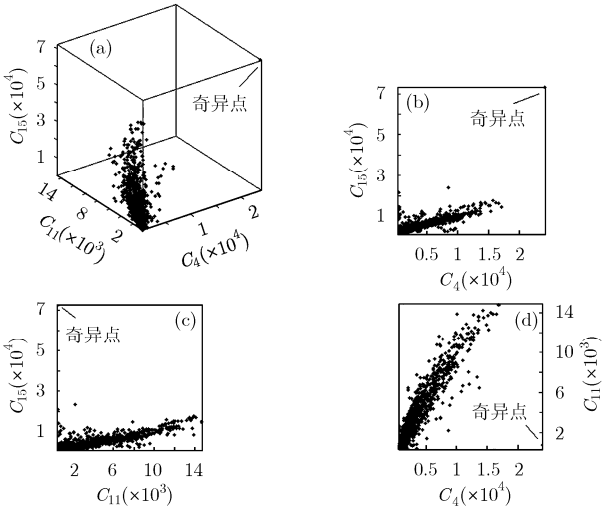

图 4 经 CPN 处理后的 3D 散点图及其各 2D 投影

6 结束语

在模式识别中，如何从维数较高的原始特征集中挑出有效特征，完成降维，是特征选择要解决的问题。常用的特征选择方法有退火法，遗传法等[1]。对于生物芯片数据的计算，找到与疾病有关基因与选择有效分类特征实际是等价的。基于此概念，作为一个有意义的推论，由于本文的CPN计算中是已经考虑了数据集的整体信息才获得IEGs集合，我们称不包含在IEGs集合中的其他基因为非不变基因non-IEGs。上面试验结果中，有41个为non-IEGs-3。对于生物医学研究，non-IEGs非常重要，包含了主要的分类/致病信息。本文实验结果意味着引起样本所患疾病的可疑基因可能就在这41个non-IEGs之中。因此，我们可使用它们作为后续分类工作的特征，所以本文工作也是一种有效的新的模式识别特征选择方法。

参 考 文 献

[1] 边肇祺，张学工等. 模式识别[M]. 北京: 清华大学出版社, 2002, 2: 205–209.

[2] Xuan J, Hoffman E, Clarke R, and Wang Y. Normalization of microarray data by iterative nonlinear regression[C]. Proc. the Fifth IEEE Symposium on Bioinformatics and Bioengineering, Minneapolis Minnesota USA, 2005: 267–270.

[3] Fujita A, Sato J R, Rodrigues L O, Ferreira1 C E, and Sogayar M C. Evaluating different methods of microarray data normalization[J]. *BMC Bioinformatics*, 2006, 7: 469.

[4] Wang Y, Lu J, Lee R, Gu Z, and Clarke R. Iterative normalization of cDNA microarray data [J]. *IEEE Trans. on Information Technology in Biomedicine*, 2002, 6(1): 29–37.

[5] Affymetrix Inc. [OL]Affymetrix technical note statistical algorithms description document. 2002, (http://www.affymetrix. com/support/technical/whitepapers/sadd_ whitepaper.pdf)

[6] Quackenbush J. Microarray data normalization and transform[J]. *Nature Genetics*, 2002, 32: 496–501.

[7] Schadt E, Li C, Eliss B, and Wong W H. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data[J]. *J. Cell. Biochem*, 2001, 84(S37): 120–125.

[8] Bolstad B M, Irizarry R A, Astrand M, and Speed T P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias [J]. *Bioinformatics*, 2003, 19(2): 185–193.

吕建平：　男，1953年生，副教授， 研究方向为生物信息、模式识别、信息安全.

Wang Yue：　男，1960年生，教授，研究方向为生物信息、模式识别.