一种基于类均值的肿瘤基因芯片数据的标准化方法

王广云[1], 邱浪波[2], 王正志[1]

(1.国防科技大学 机电工程与自动化学院, 中国湖南 长沙 410073; 2.空军工程大学 电讯工程学院, 中国陕西 西安 710077)

摘 要: 分析了当前常用的标准化方法在肿瘤基因芯片中引起错误分类的原因, 提出了一种基于类均值的标准化方法.该方法对基因表达谱进行双向标准化, 并将标准化过程与聚类过程相互缠绕, 利用聚类结果来修正参照表达水平.选取了 5 组肿瘤基因芯片数据, 用层次聚类和 K-均值聚类算法在不同的方差水平上分别对常用的标准化和基于类均值的标准化处理后的基因表达数据进行聚类分析比较.实验结果表明, 基于类均值的标准化方法能有效提高肿瘤基因表达谱聚类结果的质量.

关键词: 肿瘤基因芯片; 聚类分析; 标准化; 中心化; 相关系数

中图分类号: Q332 文献标识码: A 文章编号: 1007-7847(2007)03-0206-06

A Normalization Method for Cancer Array Data Based on Class Mean

WANG Guang-yun[1], QIU Lang-bo[2], WANG Zheng-zhi[1]

(1.College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha 410073, Hunan, China; 2.Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, Shanxi, China)

Abstract: The reasons of false classification caused by current normalization methods in cancer gene microarray are analyzed and a new normalization method based on class mean is proposed. This method normalizes gene expression profile in two directions, namely, makes normalization and cluster analysis wrapped with each other and modifies referenced expression levels using cluster results. On the different variance levels of 5 cancer gene expression profile datasets, the new method is compared with other normalization methods in hierarchical clustering and K-means clustering. The experimental results show that the proposed approach can improve the cluster results of cancer gene expression profile.

Key words: cancer array; cluster analysis; normalization; centralization; correlation coefficient

(Life Science Research, 2007, 11(3):206~211)

肿瘤基因芯片分析是当前研究的一个热点, 主要研究技术之一就是聚类分析[1,2]. 其目标是用某种相似性度量准则(如 Pearson 相关系数等)将样本或基因组织成有意义的组.对基因聚类, 有助于对基因功能、基因调控及细胞过程等进行综合研究; 对样本聚类, 可以确定和发现新的肿瘤类型, 从而对相应的诊断、治疗和预防有很大帮助. 有多种聚类算法已被成功地用于基因表达谱的聚类分析, 如层次聚类(hierarchical clustering), K-均值聚类(K-means clustering)等[3,4].

然而, 基因芯片实验中的误差来源很多, 如荧光标记效率、扫描参数的设置以及空间位置的差异等, 这些都可能对基因表达水平的测量产生影响, 从而导致较差的聚类效果. 因此, 为了消除这

收稿日期: 2007-05-09; 修回日期: 2007-07-06

作者简介: 王广云(1980-), 女, 山西运城人, 博士研究生, 从事生物信息学研究, Tel: 0731-4574991, E-mail: gfkdwgy@nudt.edu.cn.

些外界因素引起的误差，使基因表达数据能够真实地反映测量样本的生物学差异，需要对基因表达数据进行标准化处理[5]. 虽然，已有许多文献[5,7,10]介绍了当前常用的标准化方法，但还没有文献在理论上对其作用机理进行深入地阐述. 概括起来，常用的标准化方法包括零均值单位方差方法和数据中心化方法[7]. 它们都是用均值或中值对样本或基因进行标准化处理. 但是，这些方法处理后的数据不能正确反映出类别差异，在以相关系数为相似性度量准则的聚类算法中，尤其在基因表达谱中存在极端值，或者各类包含的样本或基因数量相差较大的情况下，会引起类型偏倚，从而导致样本或基因的错误分类.

针对上述问题，本文在研究聚类分析和标准化基本原理的基础上，分析了上述标准化方法引起错误分类的原因，提出了一种基于类均值的标准化方法. 该方法对基因表达谱进行双向标准化，并将标准化过程与聚类过程相互缠绕，利用聚类结果来修正基因(或样本)的参照表达水平，不但消除了芯片间差异，还突出了每个基因（或样本)在各样本(或基因)中的变异. 本文通过对 5 组寡核苷酸芯片的基因表达数据的聚类分析，验证了该方法能有效地提高聚类结果的质量.

1　聚类分析

聚类分析的基本思想是在样本或基因间定义相似性度量准则，将相似度高的样本或基因划分为一类从而确定各个样本或基因间的关系. 最常用的聚类分析方法有层次聚类(hierarchical clustering, HC)，K-均值聚类(K-means clustering, KM)等[3,4]. 这些方法都是基于个体间的相似度来进行聚类的. 因此，相似度是聚类分析的首要环节，对聚类结果有着非常重要和直接的影响.

Pearson 相关系数是最常用的相似性度量准则之一[11,12]. 它从方向上判断两个表达水平 $X=(x_1, x_2, \cdots, x_n)$ 和 $Y=(y_1, y_2, \cdots, y_n)$ 的相似程度，即

$$P=\frac{(X-\bar{X})\cdot(Y-\bar{Y})^{T}}{\left\|(X-\bar{X})\right\|\cdot\left\|(Y-\bar{Y})\right\|}=\cos\Theta, \qquad (1)$$

其中，$\bar{X}=(\bar{x})_{1\times n}$，$\bar{Y}=(\bar{y})_{1\times n}$，$\bar{x}=\sum_i x_i/n$，$\bar{y}=\sum_i y_i/n$，$\Theta$ 为向量 X 和 Y 间的夹角. P 为 1 时，X和 Y 的相似度最高，Θ 为 0°；P 为 -1 时，相反程度最高，Θ 为 180°；P 为 0 时，相关程度最低，Θ 为 90°. 可见，影响 Pearson 相关系数的是 X 和 Y 间的夹角.

2　标准化及其对聚类结果的影响

2.1　常用的标准化方法及其对聚类结果的影响

最常用的一种标准化方法是零均值单位方差，即，使每个样本或基因向量的平均值为 0，标准差为 1. 其目的是放大弱信号抑制强信号，将所有数据转换到同一个范围内. 另一种常用的标准化方法是数据的中心化，即把每个基因在各样本中的表达值减去该基因在所有样本中表达值的均值或中值来去除参照表达水平的影响，或者将各个基因在每一样本中的表达值减去该样本中所有基因表达值的均值或中值来消除芯片间的差异，使基因表达水平具有可比性. 该方法一般用于肿瘤样本的聚类或分类研究中[7].

实际上，上述两种标准化方法都有一个中心化的过程，均值和中值都是观察值“中间”位置的一种测度[6]，可以看作是对参照水平的估计. 在向量空间中，减去均值或中值就是将坐标原点平移到均值或中值所对应的点上. 零均值单位方差的标准化方法只是比数据中心化方法多了一个单位化的过程. 此过程方便比较和计算相关系数，但是，会把噪声纳入真实信号，尤其在标准差很小时会产生很大的噪声.

对样本的标准化，虽然消除了芯片间的差异，但是标准化后的值不能很好地反映各个基因在不同样本中的变异；对基因进行标准化后的值虽然突出了各个基因在不同样本中的变异，但是由于芯片间差异没有消除，各个基因在不同样本中变异的可靠性值得怀疑. 所以，只进行单向的标准化不能得到可靠的数据. 尤其值得注意的是，当对基因标准化并对样本聚类或对样本标准化并对基因聚类时[8,9,12]，由于均值和中值固有的特性[6]，会使样本（或基因）间的相似度偏离真实的相似度，从而使得聚类结果出现类型偏倚. 下面以基因芯片样本的两类别聚类为例来说明均值和中值的中心化对聚类结果的影响.

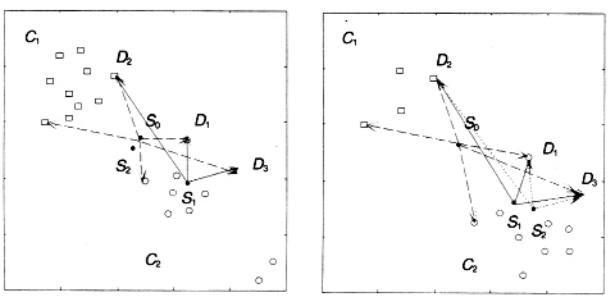
设 $A=[a_{ij}]_{m\times n}$ 为 $m\times n$ 基因表达谱矩阵，行表示基因 $g_i, i=1, \cdots, m$，列表示样本 s_j，$j=1, \cdots n$，a_{ij} 表示基因 g_i 在样本 s_j 中的表达值. 对基因中心化后，基因 g_i 的表达值为 $g_i'=g_i-[\alpha_i]_{1\times n}$，样本 s_j 的表达值为 $s_j'=s_j-\beta$，其中，α_i 表示用来中心化的值，$\beta=[\alpha_1, \cdots, \alpha_m]^T$. 使用均值中心化时，$\alpha_i=\bar{g}_i=\sum_j a_{ij}$

/n， $\beta=\bar{s}=[\bar{g}_1,\cdots,\bar{g}_m]^T$，为样本均值；使用中值中心化时，$\alpha=\hat{g}_i=\text{median}(a_{i1},\cdots,a_{in})$，$\beta=\hat{s}=[\hat{g}_1,\cdots,\hat{g}_m]^T$，为样本中值. 样本 s_i 与 s_j 经过中心化后的Pearson相关系数为

$$P=\frac{((s_i-\beta-\overline{(s_i-\beta)})^T\cdot((s_j-\beta-\overline{(s_j-\beta)})}{||(s_i-\beta-\overline{(s_i-\beta)}||\cdot||(s_j-\beta-\overline{(s_j-\beta)}||}=\cos\theta, \quad (2)$$

其中，θ 为样本中心化后向量 $(s_i-\bar{s}_i)$ 与 $(s_j-\bar{s}_j)$ 在新的坐标空间中的夹角. β取值不同，坐标原点就不同，则向量间的夹角就会发生变化，从而使样本间相似度改变.

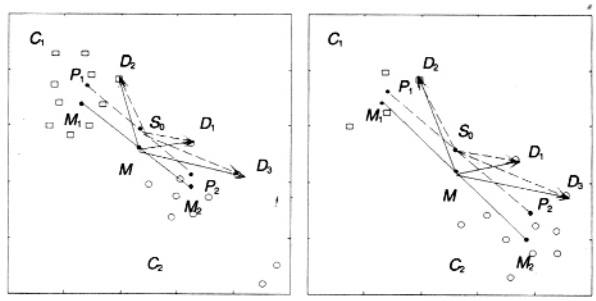
由于基因芯片具有很高的噪声，所以基因表达谱中经常出现极端值. 然而，均值对极端值很敏感. 在极端情况下，会出现只有少数几个甚至一个表达值在均值一边的情况. 此时，各样本所对应的向量 $(s_j-\bar{s}_j)$，j=1,…,n 的空间分布在二维平面上的投影如图 1a 所示. 由于 s_j 和 $(s_j-\bar{s}_j)$ 是一一对应的，为叙述方便，下面以 s_j 代替 $(s_j-\bar{s}_j)$ 进行叙述.



(a) (b)

图 1 均值和中值中心化后样本间相似度示意图

Fig.1 The similarity between samples centralized by mean and median



(a) (b)

图 2 类均值中心化后样本间相似度示意图

Fig.2 The similarity between samples centralized by class mean

图 1a 中的方块对应第一类样本点(C_1)，圆点对应第二类样本点(C_2)，且两类所包含的样本数目相等. 设 $\tilde{s}=[\tilde{g}_1,\cdots,\tilde{g}_m]^T$ 表示各基因的真实的参照表达水平，即，它对应的点 S_0 位于两类样本点的中间位置，能够很好地把两类样本分开. 经过 $\tilde{s}$ 中心化后，样本间的相似度如图 1a 中虚线箭头所示. 向量 $\overrightarrow{S_0D_1}$ 与向量 $\overrightarrow{S_0D_3}$ 间的夹角 θ_{13} 小于与向量 $\overrightarrow{S_0D_2}$ 间的夹角 θ_{12}，所以 C_2 类边缘上的点 D_1 与点 D_3 划分到 C_2 类，而点 D_2 划分到 C_1 类. 由于 C_2 类中有极端值偏离 C_2 类的其它样本点很远，所以样本均值 $\bar{s}$ 的值就偏向 C_2 类的样本表达水平，对应的点 S_1 更接近 C_2 类的样本点，如图 1a 所示. 经过 $\bar{s}$ 中心化后，样本所对应的向量如图 1 中实线箭头所示. 向量 $\overrightarrow{S_1D_1}$ 与向量 $\overrightarrow{S_1D_2}$ 间的夹角 θ_{12} 小于与向量间 $\overrightarrow{S_1D_3}$ 的夹角 θ_{13}，所以点 D_1 与点 D_2 划分到 C_1 类，而点 D_3 划分到 C_2 类. 这时就发生了错误分类. 然而，由于中值不受极端值的影响，所以样本中值 $\hat{s}$ 与 $\tilde{s}$ 比较接近，对应的点 S_2 在点 S_0 附近. 经过 $\hat{s}$ 中心化后，样本间的相似度只有略微的变化，不会影响聚类结果. 因此，当有极端存在时，均值并不是中心位置的一种好的测度.

但是，当两类所包含的样本数目不相等时，中值就会有比较大的差异. 在基因表达谱聚类分析中，两类中样本的数目一般都不会相等，还经常会出现两类中样本的数目相差较大的情况. 在这种情况下，均值和中值都会偏向数目较大的一类，使聚类结果偏向数目较小的一类. 此时，样本所对应的向量空间分布在二维平面上的投影如图 1b 所示. 图中所有标示与图 1a 相同，点线箭头表示经过 $\hat{s}$ 中心化后的样本所对应的向量. 如图 1b 所示，由于 C_2 类所含样本数目明显多于 C_1 类，点 S_1 和点 S_2 更接近 C_2 类的样本点，此时，一些原本属于 C_2 类的样本点会被划分到 C_1 类中.

2.2 基于类均值的标准化方法

为了解决上述问题，本文提出了一种基于类均值的标准化方法. 具体过程如下：

Step 1：对所有样本 s_j，j=1,…,n 进行零均值单位方差标准化；

Step 2：将样本聚为 k 类，$S_{c_{11}},\cdots,S_{c_1t_1}$ 为第一类样本 (C_1)，…，$S_{c_{k1}},\cdots,S_{c_kt_k}$ 为第 k 类样本 (C_k)，其中，$c_{11},\cdots,c_1t_1,\cdots,c_{k1},\cdots,c_kt_k=1,\cdots,n$，$t_1+\cdots t_k=n$；

Step 3: 分别计算出每一类样本的中值

$$m_1=[\text{median}(a_{1,c_1},\cdots,a_{1,ct}),\cdots,\text{median}(a_{m,c_1},\cdots,a_{m,ct})]^T$$

$$\cdots \quad , \tag{3}$$

$$mk=[\text{median}(a_{1,ck},\cdots,a_{1,ckt}),\cdots,\text{median}(a_{m,ck},\cdots,a_{m,ckt})]^T$$

以及它们的均值

$$m=(m_1+\cdots+m_k)/k, \tag{4}$$

称 m 为类均值,再将每个样本减去 m_i 对基因进行数据中心化的标准化处理,得到新的样本表达值;

Step 4: 重复 Step 2 和 Step 3,直到每类中的样本不再改变,或达到预定的迭代次数为止.

(注: 对基因的标准化也是类似的过程.)

下面以基因芯片样本的两类别聚类为例来说明该方法的有效性.

如图 2a 所示,设点 P_1 和点 P_2 分别为 C_1 类和 C_2 类的实际的类别中心,则点 S_0 位于线段 P_1P_2 的中点位置. 当样本中出现极端值时, 该方法根据第一次聚类的结果, 分别计算出 C_1 类和 C_2 类的中值 m_1 和 m_2,对应图中的点 M_1 和点 M_2. 由于点 M_2 是 C_2 类的中值点,不受极端值的影响,所以点 M_2 在点 P_2 附近; C_1 类中没有极端值, 所以点 M_1 也在点 P_1 附近. 因此, C_1 类和 C_2 类中值的均值 m 所对应的点 M 位于线段 M_1M_2 的中点位置, 并且在点 S_0 附近. 所以,经过 m 中心化后,样本间的相似度接近实际,不会影响聚类结果.

如图 2b 所示,当两类中样本的数目相差较大时, 由于本文所提出的方法先计算了每一类的中值, 所以样本数目的差异对相似度没有明显的影响. 因此,聚类结果不会受到影响.

上述过程中, m_1 和 m_2 分别是对 C_1 类和 C_2 类的类别中心的估计, 反映了每一类的基本表达水平. 经过 m 中心化后的表达值反映了每个基因在每个样本中与每个类别中心的接近程度[8], 突出了样本间的类别差异. 而且, 由于中值具有不受极端值影响的特性,所以,在初步聚类中,被错误分类的样本点对估计类别中心的影响不大. 例如,当第一次聚类时,将边缘上的点 D_1 划分到了 C_1 类中,而中值对点 D_1 的变化不敏感,只是样本数目的变化使得点 M_1 会向 C_2 类的方向稍有移动,点 M_2 会向偏离 C_1 类的方向稍有移动,但都不会偏离点 P_1 和点 P_2 很远. 这样,点 M 也不会偏离点 S_0 很远. 所以, 经过 m 中心化后再对样本聚类,将会纠正点 D_1 的错误分类.

3 实验结果

3.1 基因表达谱数据

1) 白血病数据集

选用文献[13]提供的 7 129 个白血病基因表达谱的两组数据. 第一组(Data1)有 38 个样本,包括 27 例 ALL 样本和 11 例 AML 样本; 第二组(Data2)有 34 个样本, 包括 20 例 ALL 样本和 14 例 AML 样本. 过滤掉所有表达值含有负值的基因.

还选用了文献[8] 筛选出的 50 个与 ALL 和 AML 分类紧密联系的基因(Data5),包含 25 个与 ALL 高度相关的基因,25 个与 AML 高度相关的基因. 将小于 20 的表达值改为 20.

2) 结肠癌数据集

选用文献[13]提供的 2 000 个结肠癌基因表达谱的两组数据. 第一组(Data3)有 40 个样本,包括 26 例结肠癌组织和 14 例正常组织. 第二组(Data4)有 22 个样本,包括 14 例结肠癌组织和 8 例正常组织.

3.2 结果及分析

先对所有数据进行对数变换,然后在 20 个不同方差水平上,对前 4 组数据进行特征基因筛选,每个数据集得到相应的 20 组数据. 对于 Data5 随机选取 35 个基因,使两组基因的数目有一定的差异,也得到 20 组数据. Data1-4 中行为基因列为样本, Data5 中列为基因行为样本. 对每组数据使用 4 种标准化处理方法——对列进行零均值单位方差标准化(no central, NC)、对行进行中值中心化(median central, MDC)、对行进行零均值单位方差标准化(mean central, MC)、基于类均值的标准化(class mean, CM).

为了使用已有的外部标准对聚类结果进行评估, 本文针对两类别聚类问题进行分析. 分别使用层次聚类和 K-均值聚类算法对上述数据经过 4 种预处理后得到的基因表达谱聚类. 其中, Data1-4 进行样本聚类, Data5 进行基因聚类. 表 1 和表 2 分别列出了层次聚类法和 K-均值聚类法对经过上述 4 种标准化处理后的 5 个数据集在所有方差水平上最差和最好的聚类结果. 表中数字表示聚类结果中被正确分类的样本数.

通过比较可以看出, Data1、Data2、Data3、Data5 经过 CM 标准化处理后, 在层次聚类和 K-均值聚类中的都得到了优于其它标准化处理的聚类结果,而且迭代次数不超过 6 次; Data4 无论经过

表 1 在不同方差水平和不同标准化方法下层次聚类的结果

Table 1 The results of hierarchical clustering with different variance levels and normalization methods

Dataset	No central		Median central		Mean central		Class mean		
	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Iteration times
Data1(38)	26	26	36	29	36	29	38	33	≤6
Data2(34)	21	20	32	27	32	27	33	27	≤3
Data3(40)	25	24	38	23	38	23	39	27	≤6
Data4(22)	16	11	16	16	17	16	16	16	≤3
Data5(35)	35	29	35	31	35	29	35	35	≤3

表 2 在不同方差水平和不同标准化方法下 K-均值聚类的结果

Table 2 The results of K-means clustering with different variance levels and normalization methods

Dataset	No central		Median central		Mean central		Class mean		
	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Iteration times
Data1(38)	36	28	33	28	33	29	37	32	≤6
Data2(34)	32	29	32	28	32	29	33	26	≤3
Data3(40)	38	23	36	24	36	23	37	24	≤3
Data4(22)	17	17	17	17	17	17	17	17	≤3
Data5(35)	35	29	35	31	35	29	35	35	≤6

怎样的标准化，聚类结果的正确率都不高. 这是因为 Data1-3、5 的类别差异比较显著，而 Data4 的两类样本交叉在一起，类别差异不显著. 这一点可以由 Matlab 7 中的 PCA 分析得到，此处不再赘述.

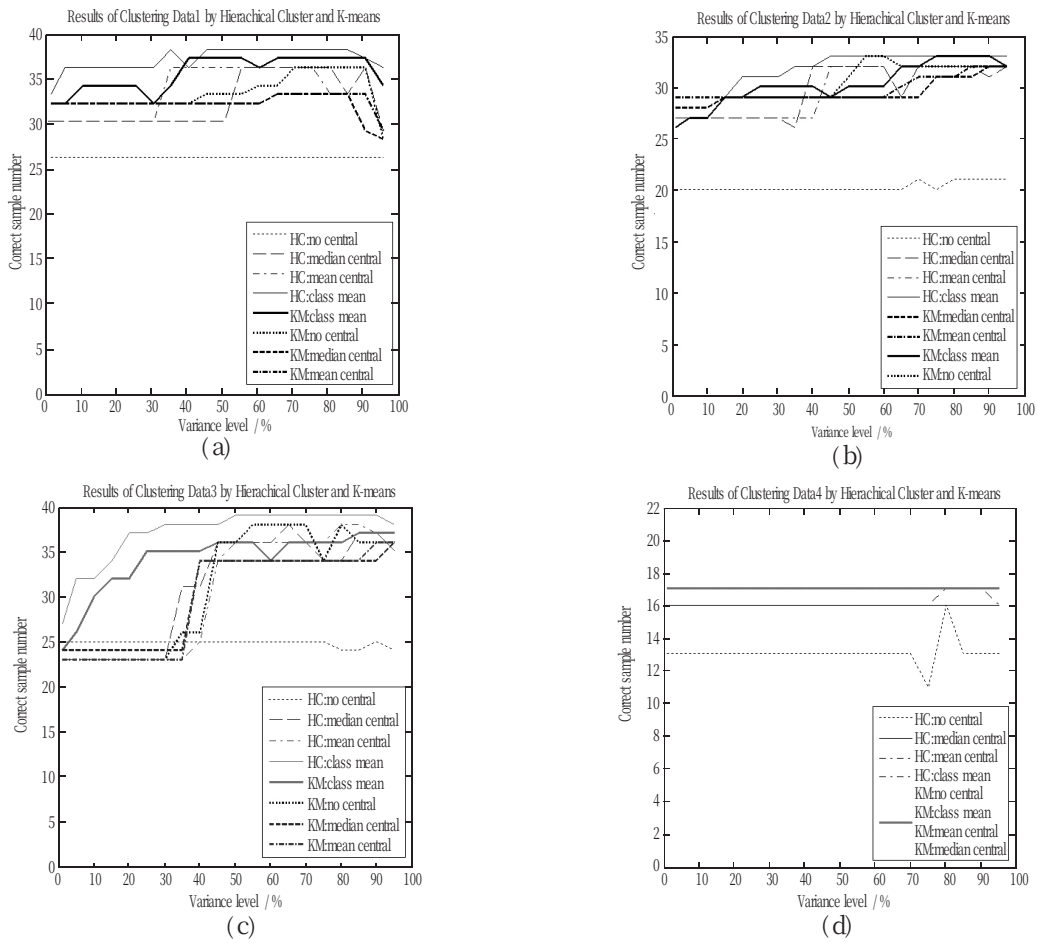


图 3 不同聚类算法中，Data1-4 在不同方差水平上聚类结果的变化曲线

Fig.3 The clustering results of Data1-4 with different variance levels and clustering methods

图 3 表示层次聚类法和 K-均值聚类法对经过上述 4 种标准化处理后的前 4 个数据集的聚类结果中被正确分类的样本数目在不同方差水平上的变化曲线(由于对 Data5 的 20 组数据是随机采样得到的没有规律性，所以不研究它的变化曲线). 其中,细线对应层次聚类法,粗线对应 K-均值聚类法. 可以看出，层次聚类法总体上要比 K-均值聚类法的结果要好. 所以，本文提出的方法更适用于层次聚类. 随着方差水平的升高，即特征基因数量的减少,无论使用哪种标准化,聚类效果都呈改善趋势,但是当基因数量太少时,又会有所下降. 从图 3 中还可以看出,Data1 和 Data3 经过 CM 标准化后的聚类结果明显优于其它标准化的聚类结果,这是由于这两个数据集中,不同类别中包含的样本数量相差较大，而且 Data1 中包含有极端值.

综上所述，本文所提出的基于类均值的标准化方法在样本聚类和基因聚类中都具有优于其它标准化方法的数据处理能力. 通过使用与聚类过程相互缠绕的迭代方法，使聚类结果得到明显改善，而且不占用时间资源. 尤其是在处理由于实验条件的限制使不同类别所包含的样本(或基因)的数目相差较大，或由于基因芯片的高噪声而使表达谱数据中包含有极端值的基因表达数据时,该方法能取得很好的效果，从而给后续的分析提供更能反映样本(或基因)间生物学差异的数据,使后续分析得到更准确的结果.

4 结论

基于类均值的标准化方法在消除芯片间差异的同时，突出了肿瘤基因在各样本中表达值与类别的相关程度，在以 Pearson 相关系数为相似度准则进行聚类时能有效的提高聚类结果的质量. 与其它标准化方法的主要区别在于，它进行双向标准化,并与聚类过程相互缠绕,所以它能够为聚类分析提供更好的数据. 本文对各种标准化方法作用机理的研究能够为研究人员提供一定的参考，帮助他们针对特定任务选择最佳的标准化处理的策略和方法.

参考文献(References):

[1] JIANG D, TANG C, ZHANG A. Cluster analysis for gene expression data: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(11): 1370-1386.

[2] AMIR B, FRIEDMAN N, YAKHINI Z. Class discovery in gene expression data[C]. Recomb, 2001: 31-33.

[3] SHERLOCK G. Analysis of large-scale gene expression data [J]. Curr Opin Immunol, 2000, 12(2): 201-205.

[4] 杨春梅,万柏坤,高晓峰. 基因表达聚类分析的现状与展望[J]. 生物化学与生物物理进展(YANG Chun-mei, WAN Bo-kun, GAO Xiao-feng. Actuality and development of the clustering technologies for gene expression[J]. Progress in Biochemistry and Biophysics), 2003, 30(6): 974-979.

[5] 贺宪民,贺佳,XIANG Zhao-ying. 基因芯片数据的标准化及分析方法[J]. 中国卫生统计(HE Xian-min, HE Jia, XIANG Zhao-ying. The normalization and analysis methods of microarray data[J]. Chinese Journal of Health Stastics), 2004, 21(2): 122-127.

[6] [美]Rosner B. 生物统计学基础(第五版)[M]. 北京:科学出版社(ROSNER B. Fundamentals of Biostatistics, 5th edition[M]. USA: Duxbury Press), 2004. 7-11.

[7] 孙啸,陆祖宏,谢建明. 生物信息学基础[M]. 北京:清华大学出版社 (SUN Xiao, LU Zu-hong, XIE Jian-ming, et al. Fundamentals of Bioinformatics[M]. Beijing: Tsinghua University Press), 2005. 288-289

[8] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(15): 531-537.

[9] ALON U, BARKAI N, NOTTERMAN D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays [J]. Proc Natl Acad Sci USA, 1999, 96(6): 6745-6750.

[10] BRAZMA A, VILO J. Gene expression data analysis[J]. FEBS Letters, 2000, 480(1): 17-24.

[11] 李瑶. 基因芯片与功能基因组[M]. 北京:化学工业出版社(LI Yao. Microarray and Functional Genomes[M]. Beijing: Chemical Industry Press), 2004. 179-180.

[12] 杨春梅,万柏坤,高晓峰. 基因聚类分析中数据预处理方式和相似度的选择[J]. 自然科学进展 (YANG Chun-mei, WAN Bo-kun, GAO Xiao-feng. The preprocessing methods of data and the selection of similarities in the clustering analysis of gene[J]. Progress in Natural Science), 2006, 16(3): 293-299.

[13] NATHALIE P, FRANK D S, JOHAN A K, et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction[J]. Bioinformatics, 2004, 20(17): 3185-3195.