基于核磁共振的代谢组学数据预处理

温锦波[1], 杨叔禹[2], 肖 娴[1], 董继扬[1*], 李学军[2], 陈 忠[1]
(1. 厦门大学物理学系, 福建 厦门 361005; 2. 厦门市第一医院, 福建 厦门, 361002)

摘要: 数据归一化是预处理中的重要组成部分, 本文采用正常大鼠与 I 型糖尿病模型大鼠的尿液核磁共振谱图作为测试数据, 研究了线归一、面归一和模归一 3 种数据预处理方法对代谢组学数据 PCA 分析结果的影响. 分析结果表明, 面归一预处理方法能够更好地在 PC 得分图上将对照组和糖尿病组的样本分开. 此外, 为了有选择性的去除代谢组学数据组中的噪声变量, 本文引入新的参数 R 来评估 PC 得分图的分类效果, 并把它引入适应度函数, 设计相应的遗传算法, 对代谢组学数据进行变量选择. 经过变量选择后, 主成分得分图上不同类别样本的可分性提高了, 而且变量数大大地减少, 更有利于特征代谢物的标记与识别.

关键词: 代谢组学; 预处理; 归一化; 变量选择
中图分类号: R 445.2 文献标识码: A 文章编号: 0438-0479(2007)06-0783-05

自从 Nicholson 于 1999 年提出代谢组学这一概念以来, 它已经成为一种有效的分析技术, 广泛应用在快速探测药物生物功能与临床诊断中[1-2]. 代谢组学是对生物体内所有代谢物进行分析, 并寻找代谢物与生理病理变化的相对关系的研究方法. 先进分析检测技术结合模式识别和专家系统等计算分析方法是代谢组学研究的基本方法.

核磁共振(NMR)由于其灵敏性、无损性并且能够根据特征峰定性的探测代谢物成分等优点已发展为代谢组学中最常用的化学分析检测技术. 基于 NMR 的代谢组学数据分析通常包括以下几个步骤[3-4]: (a) 谱图处理, 如基线校正、去偏置以及分段积分等; (b) 生成一个 m 行(每行代表一个样本) n 列(每列对应样本的一个变量) 的原始数据矩阵; (c) 数据归一化(行操作); (d) 数据标准化(列操作); (e) 数据的多元统计分析. 多元数据分析前的 4 个步骤统称为数据预处理过程[5-6]. 预处理方法有很多, 不同的预处理方法对同一个代谢组学数据集的最终分析结果产生不同的影响, 需要根据生物样品特性, 最终分析目的等因素来选择预处理方法[5-8]. 数据归一化是对数据矩阵的行操作, 意义在于消除仪器稳定度与灵敏度, 个体差异性, 以及噪声对分析结果的负面影响. 常用的归一化方法有 3 种: 线归一(Inf-Norm)、面归一(1-Norm)和模归一(2-Norm)[9-10]. 本文对比这 3 种归一化方法对 PCA 分析结果的影响, 以此探究不同归一化方法的物理特性与生物适用性.

代谢组学数据集中有些变量能够体现样本差异, 对样本分类贡献较大, 而有些变量对分类贡献较小, 或者是噪声, 这些变量会降低测量精确度或是影响一个模型的预测度[11]. 因此, 有选择的抛弃一部分变量有利于改善分类结果. 遗传算法(Genetic Algorithm, GA) 是基于自然选择与进化的优化技术, 自 20 世纪 70 年代由 Holland 提出后已经被有效用来解决变量选择问题[12]. 为了有选择性的去除代谢组学数据组中的噪声变量, 本文将 PC 得分图分类效果引入适应度函数, 设计相应的遗传算法, 在归一化之前对代谢组学数据集进行变量选择. 实验结果表明, 经过变量选择后, PC 得分图效果得到了很好的改善, 而且变量数大大的减少了, 更有利于代谢特征物的辨别.

1 实验数据采集

雄性 Wistar 大鼠 14 只, 体质量在 180~220 g 之间. 分两组, 一组 6 只, 做对照组; 另一组 8 只, 腹腔注射 STZ 造模成 I 型糖尿病大鼠. 注射后 3 d, 造模成功. 在第 4 天, 同时收集两组大鼠的 24 h 尿液. 尿液离心后立即冰冻, 在 -20℃环境保存. 两个月内实验. 配制样品时, 用磷酸盐缓冲剂(0.2 mol/L Na_2HPO_4/0.2 mol/L NaH_2PO_4, pH 7.4; 80% H_2O/20% D_2O) 1∶2 稀释尿液.

在 Varian Unity plus 500 MHz 谱仪上采集这两组大鼠尿液的 1D ^1H NMR 谱图(图 1), 用 5 mm

收稿日期: 2007-03-05
基金项目: 卫生部卫生教育联合攻关计划项目(wkj2005-2-019), 福建省自然科学基金(2007J0209), 厦门市重大疾病攻关研究基金(3502Z20051027) 资助
* 通讯作者: jydong@xmu.edu.cn

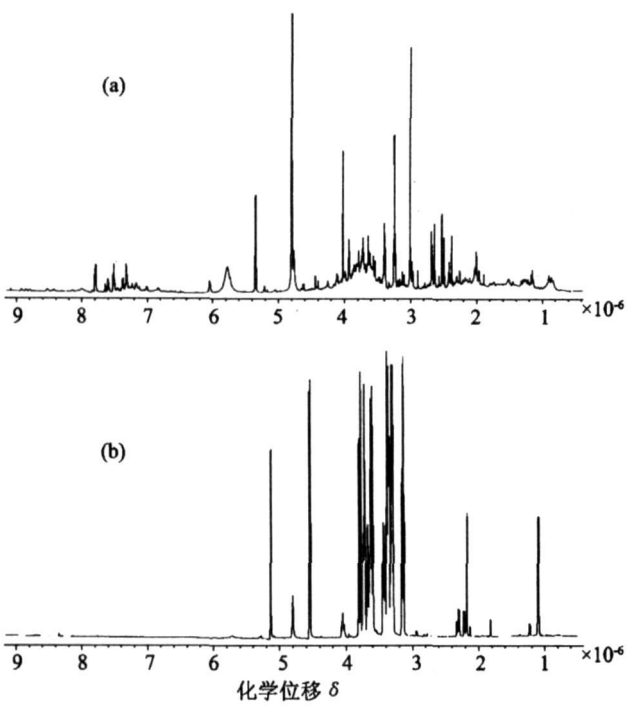

图 1 NMR 谱图

(a) 正常大鼠尿液;(b) I 型糖尿病大鼠尿液

Fig. 1 NMR spectra

HCN triple-resonance 探头, 温度 300 K, 谱宽 4.86 kHz, 样品累加次数为 128, 数据点 32 k. 采用标准的预饱和脉冲序列抑制水峰信号.

为了消除化学位移漂移的影响, 同时达到数据降维的目的, 通常采用分段积分的方法. 考虑到一般情况下化学位移漂移的大小, 积分间隔为 0.04×10^{-6}, 谱图采用内标法定标. 4.55~4.95 区间的谱峰强度设置为零, 以消除残留的水峰对分析结果的影响. 通过分段积分, 得到一个 14×243 的数据矩阵, 每行代表单个样本的所有变量, 每列代表所有样本的同一变量.

2 不同归一化方法的 PCA 分析

通常, 即使同一类样品得到的 NMR 谱图也不可能完全一致, 因为: 1) 样品之间的个体差异性, 例如浓度等引起的差异; 2) 仪器的稳定度与灵敏度, 例如发射机或是探测器的变化引起的差异; 3) 随机噪声: 采样噪声和人工噪声引起的差异. 为了尽可能地消除这些因素的影响, 归一化是必要的.

数据归一化方法的选择不仅取决于所获取的生物学信息, 而且取决于数据分析方法, 因为不同的数据分析方法对数据组的侧重面不同. 本文讨论了代谢组学中常用的 3 种数据归一化方法: 1-Norm, 2-Norm 和 Inf-Norm[9-10] (表 1).

一般地, 数据的归一化处理在数据标准化处理之前, 在基线调整、调相或移除偏置之后. 为了方便讨论, 本文中的数据标准化全部采用中心化方法.

用 PCA 对 14 个样本(6 个正常大鼠尿液样品, 8 个糖尿病大鼠尿液样品) 进行分析. 针对 3 种不同的归一化方法, 得到不同的 PC 得分图, 如图 2 所示.

从图 2 可见, 3 种归一化方法的 PCA 都能把两类样本分开, 但是每一种归一化方法对两组样品聚合效果影响却不一样. 整体上看, 正常组的聚合效果比糖尿病组好. 从生物学角度来分析, 6 个正常组样品, 代谢物成分与代谢物浓度差异很小. 对于糖尿病组, 腹腔注射 STZ 之后, 虽然都成功造模, 但是由于个体差异, 每个个体对药品的反应程度会稍有不同. 所以在 PC 得分图上, 正常组样本分布相对集中, 而糖尿病组样品相对比较分散.

为了评估 PCA 分类效果, 本文引入了一些参数来表述两类样本间的关系: (x, y) 表示样本在 PC 得分图中的坐标值, 两个类中心 (x_i, y_i), $i = 1, 2$ 的距离称为类间距(inter), 即

$$\mathrm{inter}(1,2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

其中, 类中心定义为

$$x_i = \frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right),\ y_i = \frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right),\ i = 1, 2 \quad (2)$$

n_i 为第 i 类的样本数. 第 i 类的类内距 intra(i) 定义为

表 1 常用的 3 种归一化方法描述

Tab. 1 Overview of the normalization methods used in this study

方法	描述	方程	目的
Inf-Norm	找出每个样本中最大变量, 归一化为一个常量	$x'_{i,j} = \frac{x_{i,j}}{\max_j(x_{i,j})}$	保持各个代谢成分的独立性, 消除浓度差异的影响
1-Norm	把每个样本的所有变量相加, 归一化为一个常量	$x'_{i,j} = x_{i,j} / \sum_{j=1}^{n} \lvert x_{i,j} \rvert$	显现各个数据的差异性而不是相似性, 消除浓度差异的影响
2-Norm	把每个样品的所有变量的平方相加, 归一化为一个常量	$x'_{i,j} = x_{i,j} / \sum_{j=1}^{n} x_{i,j}^2$	突出高浓度代谢成分的影响, 消除浓度差异的影响

注: $x_{i,j}$ 为样品 i 的变量, $x'_{i,j}$ 为归一化后的新变量, j 为变量编号, n 为总的变量数.

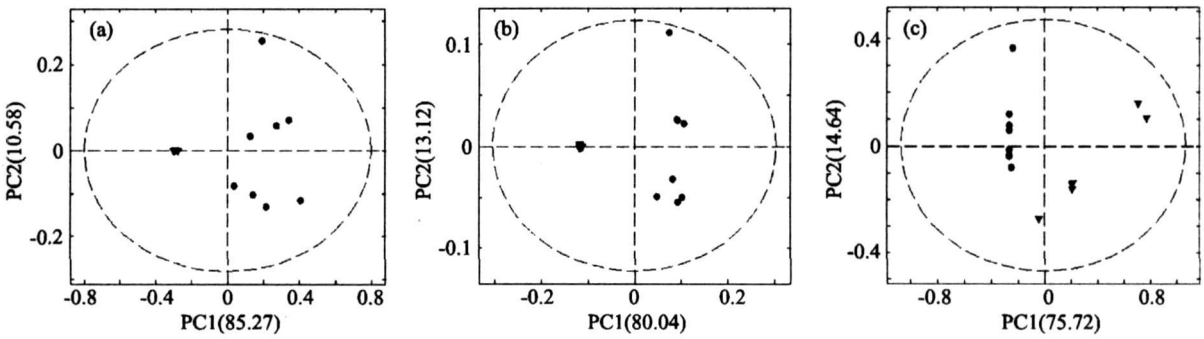

图 2　由不同归一化方法处理得到 PC 得分图

(▼) 为对照组样本点, (●) 为糖尿病组样本点. (a) Inf-Norm 方法; (b) 1-Norm 方法; (c) 2-Norm 方法

Fig. 2　PC Scoring plots of different normalization method

$$\text{intra}(i) = \frac{1}{n_i}\sum_{k=1}^{n_i}\left(\frac{1}{n_i - 1}\sum_{l=1,l\neq k}^{n_i} D_i^2(k,l)\right) \quad (3)$$

其中, $D_i(k,l)$ 表示第 i 类的第 k 个样本和第 l 个样本的距离, 即

$$D_i(k,l) = \sqrt{(x_{ik} - x_{il})^2 + (y_{ik} - y_{il})^2} \quad (4)$$

对于 PC 得分图来说, 我们希望类间距越大越好, 而类内距越小越好. 对于二分类的情况, 本文定义一个参数 R 来评估 PC 得分图的质量:

$$R = \frac{\text{inter}(1,2)}{\text{intra}(1) + \text{intra}(2)} \quad (5)$$

R 值越大, 表明分类效果越好, R 值越小, 表明分类效果越差.

为了量化地比较 3 种方法的优缺点, 根据上面定义的公式, 我们在表 2 中列出这由 3 种不同归一化方法得到的得分图参数值.

在 Inf-Norm 方法得分图中, 正常组的样品高度集中, 几乎重叠在一起, 而糖尿病组的样品在 PC1 和 PC2 上都分得很开. 我们知道, 仪器的稳定性与灵敏度体现在 PC 得分图 PC1 方向, 随机分布的噪声影响体现在 PC2 方向, 而样品差异性则同时体现在 PC1 与 PC2 方向[10]. 正常组样品的高度聚合, 说明仪器的稳定性与灵敏度的影响被消除. 由于 Inf-Norm 方法是对样本的每一个变量独立操作, 所以它能够保持各个代谢成分的独立性, 适用于生物标记物的分析. 同时, 它也保持了噪声变量的完整性, 因此 PCA 分析结果的稳定性会受到影响.

在 1-Norm 方法得分图中, 相比图 2a, 正常组的分类效果依然很好, 糖尿病组在 PC1、PC2 方向聚合度也有所提高, 尤其在 PC1 方向的聚合效果得到了明显的改善. 从表 2 中, 我们可以看出, 1-Norm 方法处理得到的得分图 R 值最高, 在得分图中也表现出最好的分类效果. 1-Norm 方法对于个体差异, 仪器稳定性与灵敏度, 以及随机噪声影响的消除都能起到作用, 它是普遍推荐的一种归一化方法.

在 2-Norm 方法得分图中, 与图 2a, b 不同, 糖尿病组样品的聚合效果是 3 种方法中最好的, 而正常组样品的聚合效果变差. 对比正常大鼠尿液与 I 型糖尿病大鼠尿液的 NMR 谱图(图 1), 发现糖尿病大鼠尿液 NMR 谱图(图 1b) 在化学位移 0.8 ~ 4.0 范围, 谱峰比正常大鼠尿液 NMR 谱图(图 1a) 高出很多, 葡萄糖、乳酸等含量明显增高, 浓度比其他代谢物高很多. 在 2-Norm 方法中加权因子是 $x_{i,j}$ 的平方, 当使用 2-Norm 方法时, 在一个样本中, 大的变量得到一个相对较大的权重. 所以在方法 2-Norm 中, 葡萄糖, 乳酸等成分的谱峰得到突显, 其它小峰变化的影响被削弱. 糖尿病组样品聚合度提高, 正常组样品聚合度相对降低. 同时从表 2 中可以看到, 图 2c 的正常组类内距与糖尿病组类内距之比等于 3.40, 而其他两种方法得到的比值为 0.06. 说明 2-Norm 方法对大信号比较敏感, 适用于需要强调高浓度代谢物变化的情况.

综上, Inf-Norm 方法能够保持各个变量的独立性, 适合用于生物标记物分析, 但是它同时保持了噪声

表 2　各个得分图的参数值

Tab. 2　Values of PCA scoring plot resulting

归一化方法	inter(1, 2)	intra(1)	intra(2)	intra(1)/intra(2)	R
Inf-Norm	109.19	2.65	45.54	0.06	2.26
1-Norm	43.42	0.76	13.39	0.06	3.07
2-Norm	177.95	79.03	23.24	3.40	1.74

的完整性,影响代谢组学数据集稳定性. 1-Norm 方法的分类效果最好,个体差异,仪器稳定度与灵敏度,和随机噪声影响的消除都能起到作用. 2-Norm 方法突出大信号的权重,对强调高浓度代谢物变化的情况比较适用.

3 变量选择对 PCA 分析结果的影响

为了去除噪声变量,进一步优化代谢组学数据集,我们在数据预处理过程中引入了变量选择方法. 遗传算法是比较常用的一种变量选择方法,它是基于自然选择与进化的优化技术,已经被用来有效地解决变量选择的问题. 在遗传算法的过程中,变量代表染色体上的基因. 通过交叉、变异和复制等遗传算子,获得具有更高适应度的新染色体,以此作为变量选择的依据.

遗传算法的关键在于适应度函数的选择. 衡量一个定量模型的好坏,常采用交叉验证均方差根(Root Mean Square Error of Cross Validation, RMSECV)表征[13],结合前面所述的 R 值,我们设计适应度函数为

$$\text{fitness}(C) = \frac{R}{\text{RMSECV}} \tag{6}$$

RMSECV 越小,R 值越大,适应度越高,模型预测效果越好. 同时,本文选取遗传迭代次数 100 为终止条件. 为了优化正常组与糖尿病组的分类,改进主成分分析的结果,在预处理过程中,我们应用遗传算法对数据组进行变量选择. 原来的 234 个变量,经变量选择后剩下 76 个变量,得到一个 14× 76 的新代谢组学 NMR 数据集. 分别应用前面的 3 种归一化方法对数据集进行 PCA 分类,经过变量选择后的数据集分类效果得到了改善(表 3).

从表 3 中可以发现,变量选择后的代谢组学数据集 PC 得分图对应的 3 种归一化方法的 R 值都有不同程度的提高. 变量选择后,2-Norm 方法与 Inf-Norm 方法的 R 值近似相等,而 1-Norm 方法的 R 值依然是 3 种方法中最大的.

表 3 变量选择前后 R 值

Tab. 3 R values with and without variable selection

方法	变量选择前	变量选择后
Inf-Norm	2.26	2.31
1-Norm	3.07	3.32
2-Norm	1.74	2.18

用遗传算法进行变量选择后,变量数从 234 降到 76 个,PC 得分图 R 值得到提高. 对比变量选择前后的得分图(图 3),我们发现正常组的样品在 PC1 上的聚合效果得到了改善,而糖尿病组中原本偏离较远的 11 号样本,在变量选择后,回到了 X 轴附近. 由于变量选择后,去除了部分噪声变量,保留了对 PCA 分类有突出贡献的变量,8 个样本在 PC2 上聚合得更好.

4 总结与讨论

本文讨论了 3 种常用的数据归一化方法对于代谢组学数据分析结果影响. 不同的归一化方法对同一数据集的侧重点不同,每一种方法都有优缺点. 1-Norm 方法对个体差异,仪器稳定度与灵敏度,和随机噪声的影响的消除都能够起到作用,表现最稳定. Inf-Norm 方法能够保持各个变量的独立性,适合用于生物标记物分析. 2-Norm 方法突出大信号的权重,适合用于需要强调高浓度代谢物变化的分析. 对归一化方法的选择取决于我们需要解决的生物问题. 数据集本身的性质以及我们所选用的分析方法. 另一方面,在数据预处

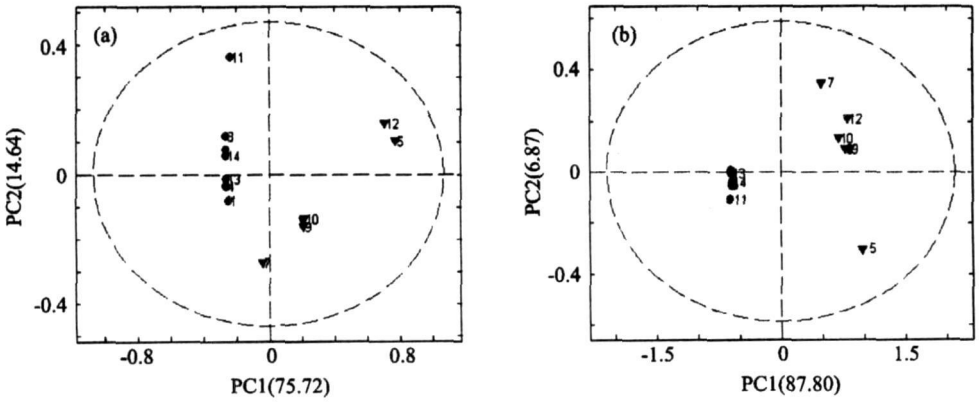

图 3 由 2-Norm 方法处理得到的 PC 得分图

(▼) 为对照组样本;(●) 为糖尿病组样本. (a) GA 之前;(b) GA 之后

Fig. 3 PC scoring plots of resulting from 2-Norm method

理过程中, 本文用 PC 得分图 *R* 值与 RMSECV 的比值做适应度函数, 设计相应的遗传算法作为变量选择方法来剔除噪声变量, 保留有用的信息, 最终达到优化 NMR 代谢组学数据集的目的, 归一化后的数据集分类效果有明显的改善. 虽然本文讨论的是基于 NMR 的代谢组学数据集优化, 但是本文所用的方法同样适用于基于质谱(MS)、气质联用技术(GC/MS)、高效液相色谱(HPLC)等的代谢组学数据集.

参考文献:

[1] Nicholls A W, Nicholson J K, Haselden J N, et al. A metabonomics approach to the investigation of drug-induced phospholipidosis: an NMR spectroscopy and pattern recognition study[J]. Biomarkers, 2000, 5: 410–423.

[2] Nicholson J K, Wilson I D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism[J]. Nature Reviews Drug Discovery, 2003, 2: 668–676.

[3] Stoyanova R, Brown T R. NMR spectral quantitation by principal component analysis[J]. NMR in Biomedicine, 2001, 14: 271–277.

[4] Lindon J C, Holmes E, Nicholson J K. Pattern recognition methods and applications in biomedical magnetic resonance[J]. Progress in NMR Spectroscopy, 2001, 39: 1–40.

[5] Craig B, Cloarec O, Holmes E, et al. Scaling and normalization effects in NMR spectroscopic metabonomic data sets[J]. Anal Chem, 2006, 78: 2262–2267.

[6] van den Berg R A, Hoefsloot H C, Westerhuis J A, et al. Centering, scaling, and transformations: improving the biological information content of metabolomics data[J]. BMC Genomics, 2006, 7: 142–156.

[7] Bro R, Smilde A K. Centering and scaling in component analysis[J]. J Chemom, 2003, 17: 16–33.

[8] Choi H K, Yoon J H. Metabolomic profiling of vitis vinifera cell suspension culture elicited with silver nitrate by H-1 NMR spectrometry and principal components analysis[J]. Process Biochemistry, 2007, 42 (2): 271–274.

[9] Webb-Robertson B J M, Lowry D F, Jarman K H, et al. A study of spectral integration and normalization in NMR-based metabonomic analyses[J]. J Phar Biom Anal, 2005, 39: 830–836.

[10] Wang Y L, Holmes E, Nicholson J K, et al. Metabonomic investigations in mice infected with schistosoma mansoni: an approach for biomarker identification[J]. Proc Natl Acad Sci USA, 2004, 101: 12676–12681.

[11] Hageman J A, van den Berg R A, Westerhuis J A, et al. Bagged K-means clustering of metabolome data[J]. C R Anal Chem, 2006, 36: 211–220.

[12] Ramadan Z, Jacobs D, Grigorov S. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms[J]. Talanta, 2006, 68: 1683–1691.

[13] Tantishaiyakul V, Worakul N, Wongpoowarak W. Prediction of solubility parameters using partial least regression[J]. International Jour of Phar, 2006, 325: 8–14.

NMR Based Metabonomic Data Preprocessing

WEN Jin-bo[1], YANG Shu-yu[2], XIAO Xian[1],
DONG Ji-yang[1*], LI Xue-jun[2], CHEN Zhong[1]

(1. Department of Physics, Xiamen University, Xiamen 361005, China;
2. Xiamen First Hospital, Xiamen 361002, China)

Abstract: Normalization is one of the most important steps of metabonomic data preprocessing. In this study, on one hand, we compared the effects of three kinds of normalization methods to the pattern recognition results in the data preprocessing, on the other hand, we evaluated evolutionary variable selection methods in improving the quality of the data clustering. Three kinds of normalization methods, i.e. Inf-Norm, 1-Norm and 2-Norm, were tested on the metabonomics data sets composed of normal and diabetes I rats' urine NMR spectra data. They were found to greatly affect the outcome of the data analysis. 1-Norm method performed better than the other two methods. Besides, parameter *R* was defined to evaluate the quality of PC scoring plot, and introduced into the fitness function of genetic algorithm (GA). The use of GA for variable selection was found to improve the data clustering quality. After GA, parts of the variables were discarding that was better to identify and recognize characteristic metabolites.

Key words: metabonomic; preprocessing; normalization; variable selection