基于偏斜 t 混合模型的流式数据自动聚类方法研究

王先文,陈 锋,程 智,杜耀华,暴洪涛,吴太虎
(军事医学科学院卫生装备研究所,天津 300161)

摘 要: 流式数据分析的主要过程是以设门的方式对样本数据中的细胞群进行类群划分.由于传统人工设门方式的缺点,提出了一种基于偏斜 t 混合模型的流式数据自动聚类方法.该方法采用有限混合模型形式,以偏斜 t 布为模型密度函数,并通过期望最大化方法估计模型参数.通过对两组不同类型实验数据进行分析,结果表明:相比于非基于模型的聚类方法,基于混合模型的聚类方法对于流式数据的分析具有更好的鲁棒性,能够降低数据中离群值对结果分析的影响;相比于高斯混合模型、偏斜正态混合模型、t 混合模型,基于偏斜 t 分布的混合模型具有更好的灵活性,不仅能够拟合流式数据中椭圆对称分布的数据,而且对于高度非对称分布数据的分析也具有很好的效果.

关键词: 混合模型; 偏斜 t 分布; 流式细胞术; EM 算法
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2014)12-2527-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.12.028

Auto Clustering Method Study of Flow Cytometry Data Based on Skew t-Mixture Models

WANG Xian-wen, CHEN Feng, CHENG Zhi, DU Yao-hua, BAO Hong-tao, WU Tai-hu
(*Institute of Medical Equipment, Academy of Military Medical Sciences, Tianjin* 300161, *China*)

Abstract: A major component of flow cytometry data analysis involves gating, which is the process of identifying homogeneous groups of cells. As manual gating is error-prone, non-reproducible, nonstandardized, and time-consuming, we propose a flexible statistical model-based clustering approach to identifying cell populations in flow cytometry data based on skew t-mixture models. This approach, which employs a finite mixture model with the density function of skew t-distribution, estimates parameters via an expectation maximization algorithm. Data analysis from two different experiments prove that the model-based clustering methods give better results in terms of robustness against outliers than non model-based clustering methods. Compared to the Gaussian mixture models, skew normal mixture models and t-mixture models, the skew t-mixture models have more flexibility in clustering symmetric data and leads to lower misclassification rates when handling highly asymmetric data.

Key words: mixture models; skew t-distribution; flow cytometry; EM algorithm

1 引言

流式细胞术(Flow CytoMetry,FCM)是一种能够精确、快速地对生物细胞的理化特性和生物学特性进行多参数定量分析及对特定细胞群分选的技术.其原理是采用微米级激光光束对经过流体动力学聚焦的细胞进行逐个激发,完整收集并记录每个细胞诱导得到的多角度散射光与多波长标记荧光信号,并通过对细胞群多光学通道数据的聚类分析实现样本的高精度定量检测.通常,单个细胞诱导得到的散射光和荧光信号以单个事件的形式被记录,所有事件汇集成被测细胞群的完整 FCM 数据.当前 FCM 数据分析的主要方法是将每个事件投影至二维或三维域中,以人工设门的方式进行分析.由于人工设门方式对数据分析人员的专业水平有较高要求,分析结果具有难以克服的主观性色彩,对实验结果的准确性与可重复性有较大影响,其事实上已形成了对流式分析技术发展的潜在威胁[1~4].

随着 FCM 相关技术研究的不断深入,近年来,人们开始探索 FCM 数据的自动分析方法,希望以此降低操作人员个人因素对实验结果的影响,推动 FCM 数据分析标准化的实现.一些自动设门的方法先后被提出[5].Demers[6]等人基于 K-means 的扩展方法能够对非对称分布的粒子进行聚类分析,但其性能相比模糊 K-means 方法[7]效果较差.Rousseeuw[8]等提出的基于启发式的模糊 K-means 方法,将每个细胞的分类进了等级化区分,该方法虽然考虑了某些分类的不确定性,但它属于一种启发

收稿日期:2014-02-21;修回日期:2014-06-26;责任编辑:覃怀银
基金项目:国家科技重大专项(No.2012ZX10004801)

式的算法并缺乏统计理论基础. Nima[9]等提出的基于 *K*-means 的快速亚群识别方法,采用变化点探测法则确定亚群数目,对于非正态分布及凹型分布的细胞群都具有较好的效果,但其对于某些高度离群细胞的处理存在一定的不确定性. Habil Zare[10]等提出的修正谱聚类方法,通过采样亚群样本数据,能够对大量数据进行快速分析,但由于该方法首先对样本数据进行采样,因此,某些信息可能在数据分析前已丢失. Yongchao[11]等后来提出的寻找局部峰方法,以 *K*-means 分析的结果作为高斯混合模型的参数,并通过混合模型分量密度函数最大梯度寻找最优 *K* 值,实现了对 FCM 数据的快速识别.

通常,细胞诱导的散射光强度与细胞的大小、形态、类型相关,而荧光强度与细胞所标记的荧光分子数量成正比,因此,流经激光照射区的具有相似内在特性的细胞的光强度会呈现一定的概率分布.最早采用基于概率分布的 FCM 数据的分析方法是高斯混合模型[12],其特点是假设 FCM 数据中的每一个事件均符合高斯分布,由于并非所有事件都满足高斯分布,且高斯分布对于离群值比较敏感,易导致模型出现过拟合现象,因此,其实用性较差. *t* 分布是一类比高斯分布具有更重尾部的分布,通过将自由度作为鲁棒性参数,可调节分布尾重,从而降低离群值对参数估计的影响. G. McLachlan 和 D Peel[13]首次采用多元 *t* 混合模型代替高斯混合模型对实验数据进行分析,结果显示 *t* 混合模型的鲁棒性更好.

偏斜 *t* 分布是基于偏斜正态分布扩展的一类分布[14,15],近年来得到巨大发展[16].由于该分布在学生 *t* 分布的基础上又增加了形状参数,能够调整分布的偏度,因此,相比于 *t* 分布其具有更好的柔韧性[17,18].基于此,本文提出了基于偏斜 *t* 分布的混合模型聚类方法.通过采用有限混合模型的形式,以概率论的方法对数据进行统计分析,实现对每一个 FCM 事件发生概率的估计,从而达到对 FCM 数据的准确聚类,提高模型分析准确度;通过采用偏斜 *t* 分布作为混合模型的密度函数,可同时调节模型中每一个分布的偏度和峰度,实现灵活地拟合样本中各亚群的分布状态,从而降低离群细胞对模型估计的影响,降低误判率.

2 方法

2.1 混合模型

通常情况,混合模型指有限混合模型,即组成模型的分量数是有限的[19].在有限混合模型中,混合密度被假定为元素密度的线性组合,即假设 $\boldsymbol{X}$ 为 p 维随机向量,则由 $\boldsymbol{X}$ 产生的多元混合模型的概率密度函数为

$$f(\boldsymbol{X};\boldsymbol{\Psi}) = \sum_{i=1}^{g}\pi_i f(\boldsymbol{X};\boldsymbol{\theta}_i) \tag{1}$$

其中 g 为混合模型的分量数,$\boldsymbol{\Psi}=(\pi_1,\cdots,\pi_{g-1},\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_g)$为未知参数矩阵,$f(\boldsymbol{X};\boldsymbol{\theta}_i)$代表第 i 个分量的密度函数,$\boldsymbol{\theta}_i$ 为第 i 个分量密度函数的未知参数向量. π_i 为混合比,表示第 i 个分量密度在混合密度中的加权因子,且满足(1)非负性:$\pi_i\geq 0$;(2)和为 1:$\sum_{i=1}^{g}\pi_i=1$.有限混合模型建模的每一个样本都被假设为由集合 $f(\boldsymbol{X};\boldsymbol{\Psi})$中随机抽取且未知的一个源 $f(\boldsymbol{X};\boldsymbol{\theta}_i)$所产生,通过对每一个源 $f(\boldsymbol{X};\boldsymbol{\theta}_i)$进行模型参数化,可建立集合 $f(\boldsymbol{X};\boldsymbol{\Psi})$的数学模型,从而辨识每一个样本是由哪个源 $f(\boldsymbol{X};\boldsymbol{\theta}_i)$产生的,并估计每一个源产生样本的概率,从而实现对所有样本的聚类分析.

2.2 多元偏斜 *t* 混合模型

根据 Sahu 等[20]提出的偏斜正态分布实现的方法,多元偏斜 *t* 分布的概率密度函数可表示为

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Omega},\boldsymbol{\delta},\upsilon)=2t_{p,\upsilon}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\upsilon)\,T_{1,\upsilon+p}\cdot\left(\frac{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\sqrt{1-\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}}}\sqrt{\frac{\upsilon+p}{\upsilon+(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}}\right) \tag{2}$$

其中,$\boldsymbol{\Sigma}=\boldsymbol{\Omega}+\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}$,$\upsilon$ 为自由度,$\boldsymbol{\delta}$ 为偏斜参数向量.由式(2)可知,当 $\boldsymbol{\delta}=\boldsymbol{0}$时,$f(\boldsymbol{x})$转化为 p 维 *t* 分布密度函数 $t_{p,\upsilon}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Omega},\upsilon)$;当 $\boldsymbol{\delta}=\boldsymbol{0}$且 υ 趋近于正无穷大时,$f(\boldsymbol{x})$转化为正态分布概率密度函数.因此,通过调节参数 $\boldsymbol{\mu},\boldsymbol{\Omega},\upsilon$ 和 $\boldsymbol{\delta}$ 值,可实现对 FCM 数据中多种亚群分布的拟合.

为使混合模型能够较好地拟合 FCM 数据中各亚群的分布状态,需要对每一个分量 $f(\boldsymbol{X};\boldsymbol{\theta}_i)$的自由参数 $\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\boldsymbol{\delta}_i$和 υ_i 值进行准确估计,这里采用极大似然估计方法.即假设 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ 为独立同分布的样本,将式(2)代入式(1)中,得到模型分量为 g 的偏斜 *t* 混合模型的似然函数:

$$L(\boldsymbol{\Psi})=\prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i f(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\boldsymbol{\delta}_i,\upsilon_i) \tag{3}$$

其中,$\boldsymbol{\Psi}=(\boldsymbol{\Psi}_1,\cdots,\boldsymbol{\Psi}_g)$为模型未知参数,可通过期望最大化算法(Expectation Maximization, EM)迭代估计 $\boldsymbol{\Psi}$ 值[21].对于偏斜 *t* 混合模型,EM 算法的计算首先需要得到完全数据的似然函数.即数据向量 $\boldsymbol{x}=(\boldsymbol{x}_1^{\mathrm{T}},\cdots,\boldsymbol{x}_n^{\mathrm{T}})^{\mathrm{T}}$认为是不完整的,因此,首先引入样本的混合分量标记向量 $\boldsymbol{z}_j=(z_{1j},\cdots,z_{gj})^{\mathrm{T}}$,且满足:$\boldsymbol{x}_j$ 属于分量 i 时,$z_{ij}=1$,否则 $z_{ij}=0$.即 z_{ij}代表样本 $\boldsymbol{x}_j$ 来自哪个分量.根据多元偏斜 *t* 分布的特性[16],再引入变量 $u_1,\cdots,u_n$ 和 $w_1,\cdots,w_n$,

$$U_j \mid w_j, z_{ij}=1 \sim \mathrm{HN}\left(0,\frac{1}{w_j}\right) \tag{4}$$

$$W_j \mid z_{ij}=1 \sim \mathrm{gamma}\left(\frac{\upsilon_i}{2},\frac{\upsilon_i}{2}\right) \tag{5}$$

其中,$\mathrm{HN}(0,\sigma^2)$代表均值为 0,方差为 σ^2 的一元半正态分布.因此,完全数据向量可表示为 $\boldsymbol{x}_c=(\boldsymbol{x}_{c1},\cdots,\boldsymbol{x}_{cn})^{\mathrm{T}}$,其中 $\boldsymbol{x}_{c1}=(\boldsymbol{x}_1^{\mathrm{T}},z_1^{\mathrm{T}},u_1,w_1)^{\mathrm{T}}\cdots\cdots\boldsymbol{x}_{cn}=(\boldsymbol{x}_n^{\mathrm{T}},\boldsymbol{z}_n^{\mathrm{T}},u_n,w_n)^{\mathrm{T}}$.通过"丢失"数据的引入[21],可得到包含变量 z_{ij}、u_j、w_j 的完全数据向量的对数似然函数:

$$\log L_c(\boldsymbol{\Psi})=\log L_{c1}(\boldsymbol{\pi})+\log L_{c2}(\boldsymbol{\theta})+\log L_{c3}(\boldsymbol{\upsilon}) \tag{6}$$

其中

$$\log L_{c1}(\boldsymbol{\pi})=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\log(\pi_i)$$

$$\log L_{c2}(\boldsymbol{\theta})=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\Big\{-\frac{1}{2}\big[p\log(2\pi)+\log|\boldsymbol{\Omega}_i|+w_j(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\delta}_iu_j)^{\mathrm{T}}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\delta}_iu_j)\big]\Big\}$$

$$\log L_{c3}(\upsilon)=\sum_{i=1}^{g}\sum_{j=1}^{n}z_{ij}\Big\{-\frac{1}{2}\big[(p-1)\log(w_j)+w_ju_j^2\big]-\frac{\upsilon_i}{2}\big[w_j-\log(\upsilon_i/2)\big]-\log\Gamma(\upsilon_i/2)+(\upsilon_i/2-1)\log(w_j)\Big\}$$

在式(6)中满足 $\boldsymbol{\pi}=(\boldsymbol{\pi}_1,\cdots,\boldsymbol{\pi}_g)^{\mathrm{T}}$,$\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_g)^{\mathrm{T}}$,$\boldsymbol{\upsilon}=(\upsilon_1,\cdots,\upsilon_g)^{\mathrm{T}}$,且 $\boldsymbol{\theta}_i$ 包含参数 $\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\boldsymbol{\delta}_i(i=1,\cdots,g)$.

EM 算法的实质是采用迭代的方法求出对数似然方程(6)的解.其计算主要分为两步:E 步用当前模型参数的估计值,对完全数据向量的对数似然函数 $\log L_c(\boldsymbol{\Psi})$关于丢失值 z_{ij}、u_j、w_j 求条件期望.即在进行第 $k+1$ 次的迭代中,E 步需要计算:

$$\tau_{ij}^{(k)}=E_{\boldsymbol{\Psi}^{(k)}}\{Z_{ij}\mid\boldsymbol{x}_j\} \tag{7}$$

$$e_{1,ij}^{(k)}=E_{\boldsymbol{\Psi}^{(k)}}\{W_j\mid\boldsymbol{x}_j,z_{ij}=1\} \tag{8}$$

$$e_{2,ij}^{(k)}=E_{\boldsymbol{\Psi}^{(k)}}\{W_jU_j\mid\boldsymbol{x}_j,z_{ij}=1\} \tag{9}$$

$$e_{3,ij}^{(k)}=E_{\boldsymbol{\Psi}^{(k)}}\{W_jU_j^2\mid\boldsymbol{x}_j,z_{ij}=1\} \tag{10}$$

$$e_{4,ij}^{(k)}=E_{\boldsymbol{\Psi}^{(k)}}\{\log(W_j)\mid\boldsymbol{x}_j,z_{ij}=1\} \tag{11}$$

其中

$$\tau_{ij}^{(k)}=\frac{\pi_i^{(k)}f(\boldsymbol{x}_j;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Omega}_i^{(k)},\boldsymbol{\delta}_i^{(k)},\upsilon_i^{(k)})}{\sum_{i=1}^{g}\pi_i^{(k)}f(\boldsymbol{x}_j;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Omega}_i^{(k)},\boldsymbol{\delta}_i^{(k)},\upsilon_i^{(k)})}$$

表示观察样本 $\boldsymbol{x}_j$ 属于第 i 个分量的后验概率,即样本 $\boldsymbol{x}_j$ 来自于混合模型中源 i 的概率,$\tau_{ij}^{(k)}$值的估计实现了模型对 FCM 数据中每一个事件发生概率的估计.

M 步对 E 步条件期望后的对数似然函数求极大似然估计,重新估计模型参数,即求

$$\pi_i^{(k+1)}=\frac{1}{n}\sum_{j=1}^{n}\tau_{ij}^{(k)} \tag{12}$$

$$\boldsymbol{\mu}_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k)}(\boldsymbol{x}_je_{1,ij}^{(k)}-\boldsymbol{\delta}_i^{(k)}e_{2,ij}^{(k)})\Big/\sum_{j=1}^{n}\tau_{ij}^{(k)}e_{1,ij}^{(k)} \tag{13}$$

$$\boldsymbol{\Omega}_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k)}\Big\{e_{1,ij}^{(k)}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(k)})(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(k)})^{\mathrm{T}}-e_{2,ij}^{(k)}\boldsymbol{\delta}_i^{(k)}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(k)})^{\mathrm{T}}-(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(k)})\boldsymbol{\delta}_i^{(k)}e_{2,ij}^{(k)\mathrm{T}}+e_{3,ij}^{(k)}\boldsymbol{\delta}_i^{(k)}\boldsymbol{\delta}_i^{(k)\mathrm{T}}\Big\}\Big/\sum_{j=1}^{n}\tau_{ij}^{(k)} \tag{14}$$

$$\boldsymbol{\delta}_i^{(k+1)}=\frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}e_{2,ij}^{(k)}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(k)})}{\sum_{j=1}^{n}\tau_{ij}^{(k)}e_{3,ij}^{(k)}} \tag{15}$$

$$\sum_{j=1}^{n}\tau_{ij}^{(k)}[\log(\upsilon_i^{(k+1)}/2)-\psi(\upsilon_i^{(k+1)}/2)+1]+\sum_{j=1}^{n}\tau_{ij}^{(k)}(e_{4,ij}^{(k)}-e_{1,ij}^{(k)})=0 \tag{16}$$

其中 $\psi(s)=\{\partial\Gamma(s)/\partial s\}/\Gamma(s)$为伽马函数的导数.通过不停地迭代 E 步和 M 步,使似然函数的差 $L(\boldsymbol{\Psi}^{(k+1)})-L(\boldsymbol{\Psi}^{(k)})$变很小(小于某设定值),收敛时的 $\boldsymbol{\Psi}$ 值即为模型参数的估计值.

2.3　初始值的确定

由以上对偏斜 t 混合模型的 EM 算法分析可知,模型参数的估计对于初始值比较敏感,因此,需要对初值 $\boldsymbol{\Psi}$ 值进行设定.本文采用 K-means 方法,其具体实施过程如下:

(1)根据混合模型分量数 g 采用 K-means 方法对数据向量 $\boldsymbol{x}$ 进行初始聚类;

(2)根据 K-means 聚类结果计算分量标记向量 $\boldsymbol{z}_j^{(0)}=\{z_{ij}^{(0)}\}_{i=1}^{g}$;

(3)初始化混合比 $\pi_i^{(0)}=\frac{1}{n}\sum_{j=1}^{n}z_{ij}^{(0)}$;

(4)初始化 $\boldsymbol{\mu}_i,\boldsymbol{\Omega}_i,\boldsymbol{\delta}_i$ 和 υ_i,其具体计算如下:

$$\boldsymbol{\mu}_i^{(0)}=\frac{\sum_{j=1}^{n}z_{ij}^{(0)}\boldsymbol{x}_j}{\sum_{j=1}^{n}z_{ij}^{(0)}}$$

$$\boldsymbol{\Omega}_i^{(0)}=\frac{\sum_{j=1}^{n}z_{ij}^{(0)}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(0)})(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(0)})^{\mathrm{T}}}{\sum_{j=1}^{n}z_{ij}^{(0)}}$$

$$\boldsymbol{\delta}_i^{(0)}=\mathrm{sgn}\Big\{\sum_{j=1}^{n}[z_{ij}^{(0)}(\boldsymbol{x}_j-\boldsymbol{\mu}_i^{(0)})]^3\Big\}$$

$$\upsilon_i^{(0)}=4$$

良好的初值有助于避免模型收敛到局部最优,并快速收敛到最优解,从而提高模型计算效率.

2.4　类群数目的确定

根据以上分析,模型的建立及模型参数的估计首先需要确定类群数目,即模型参数 g,本文采用贝叶斯信息准则(Bayesian Information Criterion,BIC)方法[22,23],该准则由 Schwarz 在 Akaike 信息准则[24](Akaike Information Criterion,AIC)的基础上提出,基于随机建模的思想,借助信息论方法,通过使准则达到最小值确定模型阶次.在偏斜 t 混合模型中,其定义如式:

$$\mathrm{BIC}=-2\log L+k\log n \tag{17}$$

其中 L 是式(3)中混合模型极大似然估计的对数似然值,k 为混合模型的独立参数,n 表示样本大小,即样本容量.通过计算每一个分量值 k(取值范围为 1 ~ g)对应的 BIC 值,选择其中最小 BIC 对应的 k 值,即为模型

分量数.

3 结果及分析

为验证本文设计的方法对于 FCM 数据的自动分析能力,将模型应用于流式实验结果的分析中.实验一的数据来自于某病人的外周血样本,其包含 14656 个细胞以及 3 种标记分子,即 CD3,CD8 和 CD4.实验以 $CD4^+$ 细胞为目标细胞进行分析说明.通常 FCM 数据分析的第一步是对前向散射光(Forward Light Scatter,FSC)和侧向散射光(Sideward Light Scatter,SSC)数据设门分析,通过这两维数据的分析能够检测细胞的相对形态属性(细胞大小和细胞形状),从而区分细胞基本类型(如单核细胞和淋巴细胞)或排除死细胞和杂质.

图 1 显示的是实验操作人员采用 FloMax 软件对实验数据分析的结果,可以看出,图中门 $R1 \sim R4$ 分别对应的为淋巴细胞亚群、单核细胞群、粒细胞群及破碎的红细胞和血小板.据图可知,门 $R1$ 和 $R2$ 对应的数据基本成椭圆对称分布,而门 $R3$ 和 $R4$ 对应的数据呈非对称分布,且门 $R4$ 对应的数据,其分布明显具有较长尾部.

图 2 显示的是 g 从 1 变化到 10 时模型 BIC 值的变化,从图中可以看出,在 g 等于 3 变化到 4 的过程中,BIC 值变化较大,而之后几乎处于相对稳定的状态,因此,模型分量数确定为 4 是最合适的.与专家分析的结果对比,采用 BIC 方法计算的模型分量数是一致的.

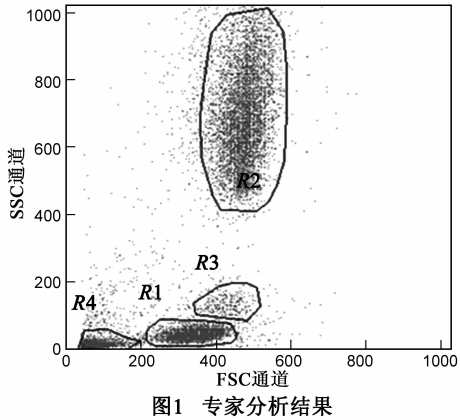


图1 专家分析结果

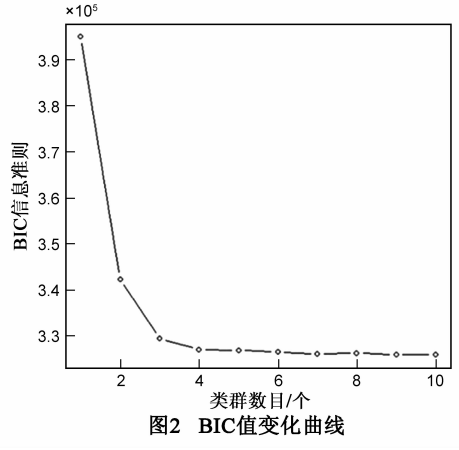
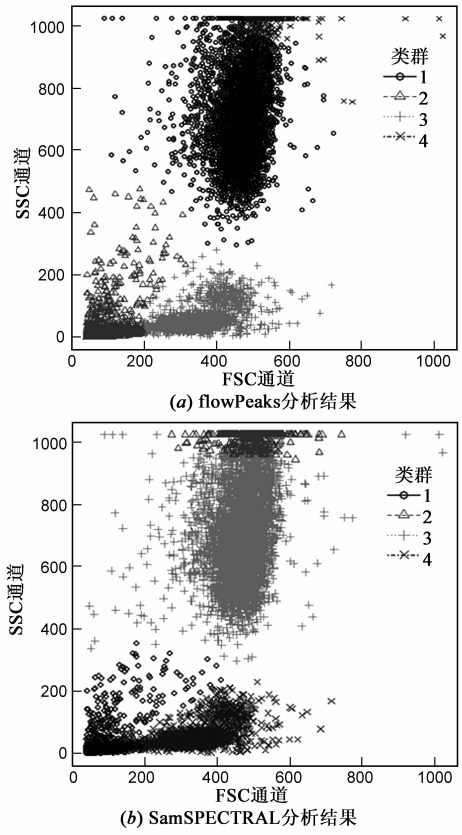


图2 BIC值变化曲线

在确定模型分量数后,接下来是样本数据的自动设门分析.为验证混合模型分析方法对 FCM 数据分析的优效性,实验对比了非基于混合模型的方法:基于 K-means 的 flowPeaks[11] 方法和基于谱聚类的 SamSPECTRAL[10]方法.图 3(a)、图 3(b)分别为两种方法分析的结果,对比图 1 的结果,虽然 flowPeaks 和 SamSPECTRAL 均能够识别正确的亚群数目,但由于数据中极小部分野值的影响,两种方法分类结果与专家分析结果误差较大.



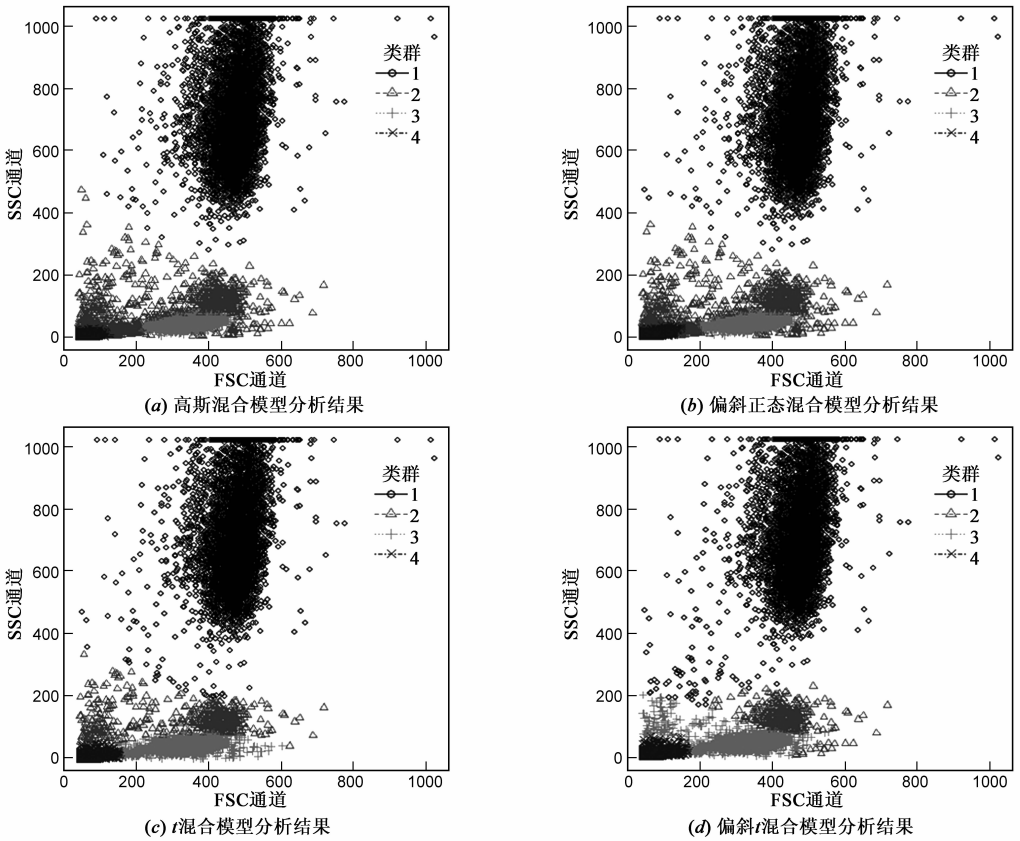
(a) flowPeaks分析结果

(b) SamSPECTRAL分析结果

图3 两种方法分析结果对比

同时,实验还对比了目前主要的混合模型方法:高斯混合模型、偏斜正态混合模型、t 混合模型及本文设计的偏斜 t 混合模型,其分析结果如图 4(a)~4(d).与专家分析的结果对比来看,高斯混合模型对于基本呈椭圆对称分布的数据拟合性较好,但由于高斯分布对于离群值较为敏感,对于非对称分布的数据其拟合结果较差.相比于高斯分布,由于偏斜正态分布和 t 分布分别增加了偏度和尾重调节参数,偏斜正态混合模型与 t 混合模型对非对称分布的数据拟合性相对较好.图 4(d)显示的是偏斜 t 混合模型分析的结果,由于偏斜 t 分布同时具有尾重和偏度调节参数,使得其能够更加

灵活地拟合含有高度非对称分布的数据.



(*a*) 高斯混合模型分析结果　(*b*) 偏斜正态混合模型分析结果

(*c*) *t*混合模型分析结果　(*d*) 偏斜*t*混合模型分析结果

图4　四种混合模型分析结果对比

图 5 显示的是去除离群值后的偏斜 t 混合模型分析的结果(在这里,以 90% 作为正常值和离群值的分界点),"*"标记的是离群细胞.为了获得 $CD4^+$ 细胞的数量,将结果中数据的 CD3 和 CD8 通道的荧光数据再进行投影,然后采用各方法进行分析.图 6 为专家分析的结果,QA2 区域表示的为 $CD4^+$ 细胞.图 7(*a*)、(*b*)分别为采用偏斜 t 混合模型方法的 BIC 值变化曲线和聚类结果.由图可知,偏斜 t 混合模型很好地降低了离群细胞(图 7(*b*)中边缘部分细胞群)对结果分析的影响,图 7(*b*)中"+"标记的为 $CD4^+$ 细胞.

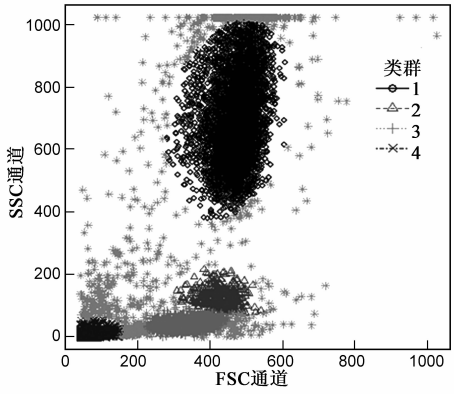


图5　去除离群值后的聚类结果

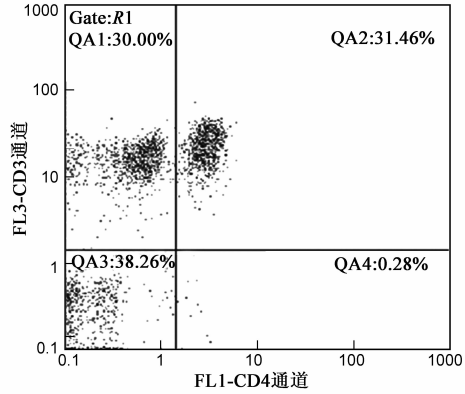
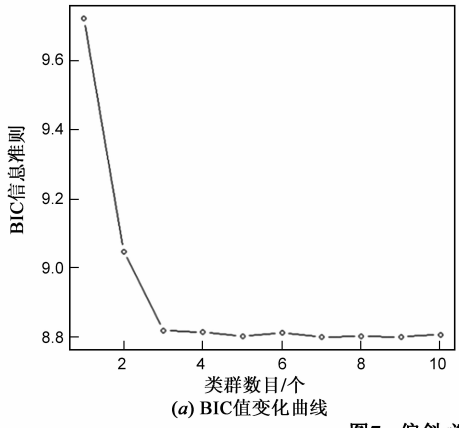
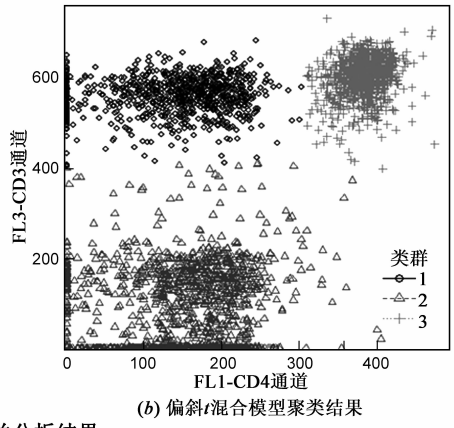


图6　专家分析的结果

对比以上几种分析方法,并以专家分析的结果作为基准,计算各方法的误判率(Misclassification Rates, MR),即细胞被分配到错误类群的概率,其计算时首先将破碎的红细胞和杂质数据移除,计算结果如表 1.由结果可得,相比于 flowPeaks 和 SamSPECTRAL,偏斜 t 混合模型的分析结果与专家分析的结果最相近,其误判率最低(0.06442),而高斯混合模型是混合模型中结果最差的(MR = 0.28524).



(*a*) BIC值变化曲线



(*b*) 偏斜*t*混合模型聚类结果

图7 偏斜*t*混合模型相关分析结果

表 1 不同分析方法的误判率比较

	分析方法					
	flowPeaks	SamSPECTRAL	高斯混合模型	偏斜正态混合模型	*t* 混合模型	偏斜 *t* 混合模型
MR	0.28253	0.30698	0.28524	0.21151	0.13431	0.06442

为进一步验证本文设计的算法对于流式数据分析的优效性,实验二对另一组粒子总数为 10000 并含有 FSC、SSC、FL1-FDA、FL2-PI 的实验数据进行了分析.该数据来自于酵母菌活性检测实验,数据分析的第一步是通过 FSC 和 SSC 组成的散点图排除样本中的少部分杂质,找到酵母菌群.由于样本中只含有酵母菌,因此 FSC 和 SSC 散点图中只有一个类群分布,在此不对其进行详细聚类分析.排除杂质后,接下来是将选择的酵母类群在 FL1-FDA 和 FL2-PI 维度进行分析,找出各种状态酵母菌群的绝对数量及所占比例.如图 8 是专家采用 FloMax 软件设门分析的结果,其中 $Q1$ 区域代表的是死菌,其数量极少,约占总数的 0.95%.$Q2$ 区域代表的是处于死活状态之间的菌(细胞膜已破裂,PI 已进入细胞),$Q3$ 区域代表的是细胞碎片,$Q4$ 区域代表的是活菌.由图可知,该数据中包含了长拖尾的亚群数据($Q2$ 区域).图 9 是采用 BIC 方法计算结果.

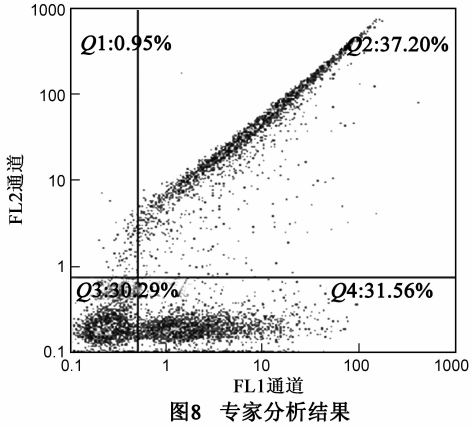


图8 专家分析结果

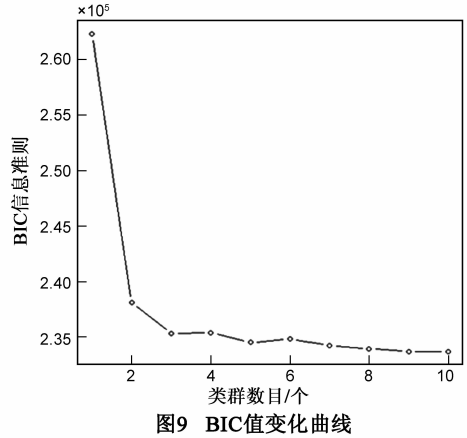
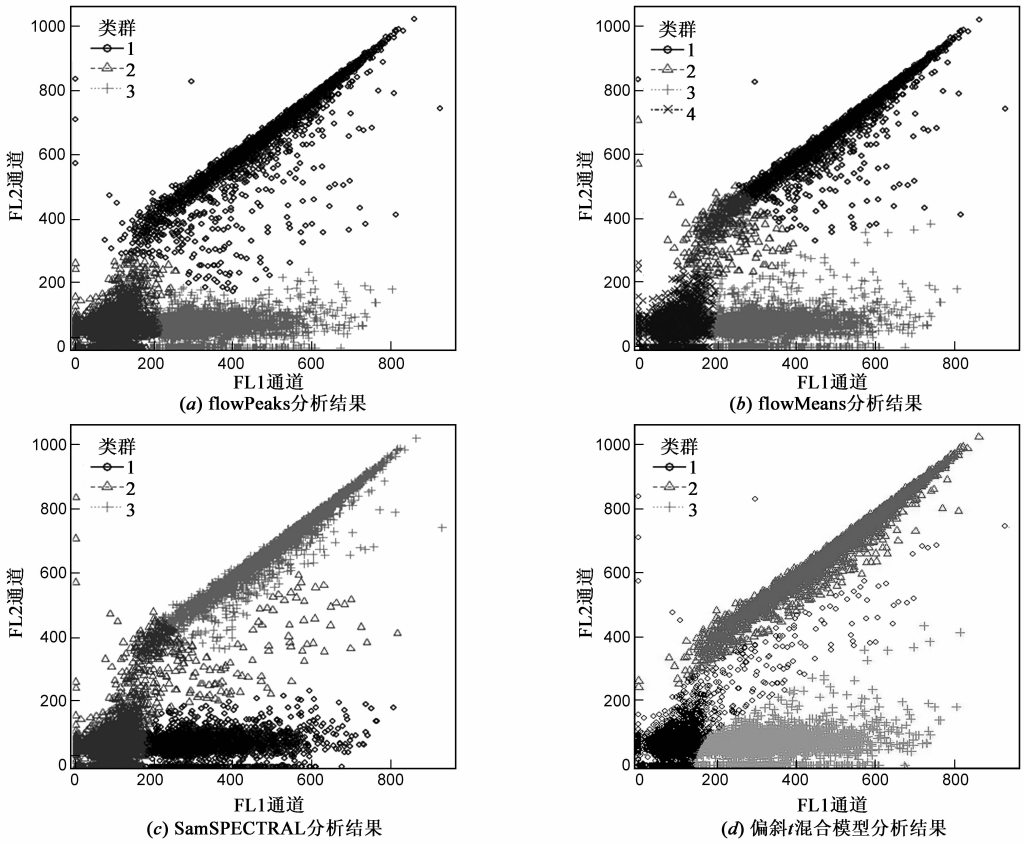


图9 BIC值变化曲线

实验采用了 flowPeaks、flowMeans 和 SamSPECTRAL 及偏斜 *t* 混合模型对数据进行分析,相应分析结果如图 10(a)~(d).由结果可知,对于 $Q4$ 区域活酵母细胞群的识别,四种方法聚类结果基本一致.但对于 $Q2$ 区域菌群的识别,flowMeans 分析结果出现了过拟合现象,而 SamSPECTRAL 虽然类群数目识别正确,但聚类结果与专家分析结果有较大差别,flowPeaks 与本文设计的偏斜 *t* 混合模型分析结果均较好,且偏斜 *t* 混合模型最接近专家分析结果.



(*a*) flowPeaks分析结果　(*b*) flowMeans分析结果

(*c*) SamSPECTRAL分析结果　(*d*) 偏斜*t*混合模型分析结果

图10　四种方法分析结果对比

4　讨论

为验证基于偏斜 t 混合模型对 FCM 数据自动分析的能力以及相比于其他方法的优势,本文选取了两种不同类型且具有代表性的实验数据进行分析.数据中既包含符合对称分布的数据,还包含具有长拖尾高度非对称分布的数据,实验数据的分析能够有力的说明偏斜 t 混合模型对于 FCM 数据分析的有效性.在结果分析中,以 90% 作为正常值和离群值的分界点是根据经验来定的,在实际的数据分析中可进行修改.其中,分界点是根据样本细胞在亚群分布中所对应的概率值大小来进行划分的.

在有限混合模型中,模型分量数和模型分量密度族决定了混合模型的结构和效能[19].对于模型分量数的估计,常用的方法是交叉验证和各种信息准则,本文选用的是 BIC 准则.该准则对于混合模型非常有效[25,26],在计算极大似然估计的时候一并计算出来.当 FCM 数据较大的时候,其优势更加明显.相对于 AIC 准则,BIC 准则将样本容量引入到实际参数估计中,从而避免因样本量太大而淹没独立模型参数对参数估计的影响[22].

相比于其他聚类方法,基于混合模型的聚类方法通过已知模型密度函数对数据中的类群进行拟合,其不仅能够将数据分成预定的类,还能够根据密度函数精确地计算出数据中每一个事件所属每一个类群的概率[27].在实际的 FCM 数据分析中,后验概率 $\tau_{ij}^{(k)}$ 可理解为粒子 j 属于亚群 i 的可能性.因此,当为某个微粒指定类型时,应选择 $\tau_{ij}^{(k)}$ 值最大的那个 i 值作为微粒 j 的所属类.与混合模型不同,flowPeaks 以 K-means 算法为核心,采用分量密度函数最大梯度寻找局部峰个数 k,并根据两峰间距大小合并获得最终结果.由于以粒子之间的距离为分类依据,当数据中存在小部分高度集中野值时,野值可能连同部分正常值被划分为另一类群,从而对结果产生影响,导致模型出现过拟合现象,如图 3(a).与 flowPeaks 不同,SamSPECTRAL 是一种基于谱聚类的方法,通过采用数据预抽样的方式,能够避免传统谱聚类方法计算量大的缺点,但是,由于数据分析前首先进行了抽样,当数据中存在数量较小的类群时,该类群可能在抽样后变得难以识别,从而导致方法对于该部分类群的错误判别,如图 3(b)中粒细胞群.同样的,由于密度较低且紧邻高密度类群,flowPeaks 也很难识别图 3(a)中粒细胞亚群并得到正确的分析结果.

从实验分析的结果来看,偏斜 t 混合模型对于 FCM 数据分析的鲁棒性比其他混合模型较好.在本文设计

的模型中,首先通过 BIC 方法求取最佳模型分量数,而不是先排除离群值后再进行模型分量数的确定,这样做是因为偏斜 t 混合模型具有更强的稳健性(图 4(d)和图 10(d)),能够适应更长拖尾的野值.对于高斯混合模型,由于高斯密度函数的分布特性,当数据中出现非对称分布的类群时,其无法正确拟合数据[28].并且,当数据中野值太多时,由于需要容纳这些野值,高斯混合模型也可能出现过拟合或分类错误的情况(图 4(a)).相比于高斯混合模型,t 混合模型和偏斜正态混合模型虽然在鲁棒性方面有所加强,但对于含有高度非对称分布的流式数据,其结果也难以达到分析要求,如图 4(b)、图 4(c).

在 EM 算法的计算过程中,模型初始值对 EM 计算影响非常大,一个不好的初始条件可能导致漫长的收敛过程或者导致一个局部收敛值.在实际 FCM 数据的分析中,本文采用了 K-means 方法对模型参数进行初始化.然而,更好的初始值设置方法或许需要纳入一些专家知识,或者需要一些更复杂的数据集.另外,由 EM 算法的实施过程可知,本文设计的算法的计算时间复杂度为 $O(d^4nktm)$,其中 d 代表分析 FCM 数据的维度,n 为 FCM 数据中的粒子总数,k 为类群数目,t 为 K-means 初始化模型参数的迭代次数,m 为找到最佳类群数目的时间.因此,当分析的 FCM 数据维度较多时,可采用将数据进行投影的方式进行聚类分析.在本文实际的数据分析中,以主频为 3.2GHz,内存为 4GB 的计算机为平台,两次实验分别用时约 42s 和 29s.其中,由于模型参数自由度 $v_i^{(k+1)}$ 的解不存在封闭的形式(可采用式(16)右边的估计值或将 $e_{4,ij}^{(k)}$ 用无限级数进行展开的方法来进行求解),为减少计算时间,实验中采用令 $\upsilon_i = 4$ ($i = 1, \cdots, g$)的方法.

5 结论

由于 FCM 的快速发展,传统人工数据分析的方法已难以满足多维度、高通量数据分析的需求[29].本文基于 FCM 数据分布特点,提出了基于偏斜 t 分布的混合模型聚类方法.实验结果表明,相比于非基于模型的方法和其他混合模型,本文提出的方法具有很好的鲁棒性,能够拟合 FCM 数据中各种分布状态的数据,尤其对于高度非对称分布数据的分析,具有较低的误判率,是目前实现 FCM 数据准确分析较理想的方法.

参考文献

[1] Satoh C, Dan K, Yamashita T, et al. Flow cytometric parameters with little interexaminer variability for diagnosing low-grade myelodysplastic syndromes [J]. Leukemia Research, 2008, 32(5): 699 – 707.

[2] Gratama JW, Kraan J, Keeney M, et al. Reduction of variation in T-cell subset enumeration among 55 laboratories using single-platform, three or four-color flow cytometry based on CD45 and SSC-based gating of lymphocytes[J]. Cytometry, 2002, 50(2): 92 – 101.

[3] Van Blerk M, Bernier M, Bossuyt X, et al. National external quality assessment scheme for lymphocyte immunophenotyping in Belgium[J]. Clinical Chemistry and Laboratory Medicine, 2003, 41(3): 323 – 330.

[4] Hahne F, Khodabakhshi AH, Bashashati A, et al. Per-channel basis normalization methods for flow cytometry data[J]. Cytometry Part A, 2010, 77(2): 121 – 131.

[5] AliBashashati, Ryan R Brinkman. A survey of flow cytometry data analysis methods [J/OL]. Advance in Bioinformatics, 2009, 2009: Article ID 584603, doi: 10.1155/2009/584603.

[6] Demers S, Kim J, Legendre P, et al. Analyzing multivariate flow cytometric data in aquatic sciences[J]. Cytometry, 1992, 13(3): 291 – 298.

[7] Wilkins MF, Hardy SA, Boddy L, et al. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data[J]. Cytometry, 2001, 44(3): 210 – 217.

[8] Rousseeuw P J, Kaufman L, Trauwaert E. Fuzzy clustering using scatter matrices[J]. Comput Statist Data Anal, 1996, 23(1): 135 – 151.

[9] Nima Aghaeepour, Radina Nikolic, Holger H Hoos, el at. Rapid cell population identification in flow cytometry data[J]. Cytometry Part A, 2011, 79(1): 6 – 13.

[10] Habil Zare1, Parisa Shooshtari1, Arvind Gupta, et al. Data reduction for spectral clustering to analyze high throughput flow cytometry data[J/OL]. BMC Bioinformatics, 2010, 11: 403. DOI: 10.1186/1471 – 2105 – 11 – 403.

[11] Yongchao Ge, Stuart C Sealfon. FlowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding[J]. Bioinformatics, 2012, 28(15): 2052 – 2058.

[12] Boedigheimer MJ, Ferbas J. Mixture modeling approach to flow cytometry data[J]. Cytometry Part A, 2008, 73(5): 421 – 429.

[13] Geoffrey J Mc Lachlan, D Peel. Robust cluster analysis via mixtures of multivariate t-distributions[J/OL]. Advances in Pattern Recognition of Lecture Notes in Computer Science, 1998, 1451: 658 – 666. DOI: 10.1007/BFb0033290.

[14] A Azzalini. A class of distributions which includes the normal ones[J]. Scandinavian Journal of Statistics, 1985, 12(2): 171 – 178.

[15] Tsung I Lin, Jack C Lee, Shu Y Yen. Finite mixture modeling using the skew normal distribution[J]. Statistica Sinica, 2007, 17(3): 909927.

[16] Tsung I Lin, Jack C Lee, Wan J Hsieh. Robust mixture model-

ing using the skew-t distribution[J]. Statistics and Computing, 2007,17(2):81 – 92.
[17] A Azzalini, A Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2003,65(2):367 – 389,
[18] A K Gupta. Multivariate skew t-distribution [J]. Statistics, 2003,37(4):359 – 363.
[19] G J McLachlan, D Peel. Finite Mixture Models [M]. New York: Wiley,2000.23 – 30.
[20] Sujit K Sahu, Dipak K Dey, Marcia D Branco. A new class of multivariate skew distributions with applications to Bayesian regression models [J]. Canadian Journal of Statistics,2003,31 (2):129 – 150.
[21] A P Dempster, N M Laird, D B Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion) [J]. Journal of the Royal Statistical Society, Series B. 1977,39 (1):1 – 38.
[22] Schwarz G. Estimating the dimension of a model[J]. The Annals of Statistics,1978,6(2):461 – 464.
[23] 刘伟峰,杨爱兰.基于 BIC 准则和 Gibbs 采样的有限混合模型无监督学习算法[J].电子学报,2011,39 (3A):134 – 139.
LIU Wei-feng, YANG Ai-lan. Unsupervised Learning for finite mixture models based on BIC criterion and Gibbs sampling[J]. Acta Electronica Sinica,2011,39(3A):134 – 139. (in Chinese)
[24] Akaike H. A new look at the statistical identification model [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716 – 723.
[25] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis[J]. The Computer Journal,1998,41(8):578 – 588.
[26] Chris Fraley, Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation [J]. Journal of American Statistical Association,2002,97(458):611 – 631.
[27] Haixian Wang, Zilan Hu. On EM estimation for mixture of multivariate t-distributions[J]. Neural Process Letters, 2009, 30(3):243 – 256.
[28] Wenlong Huang, Licheng Jiao. Unsupervised texture segmentation based on artificial immune Gaussian mixture models network[J]. Chinese Journal of Electronics,2008,17(2):301 – 304.
[29] J Paul Robinson, Bartek Rajwa, Valery Patsekin, et al. Computational analysis of highthroughput flow cytometry data[J]. Expert Opinion on Drug Discovery,2012,7(8):679 – 693.

作者简介

王先文　男,1987 年 1 月出生于四川省绵阳市,现为军事医学科学院卫生装备研究所博士研究生,主要研究方向为模式识别与人工智能.
E-mail:wangxianwen_work@126.com

吴太虎(通讯作者)　男,1962 年生,山西沁县人,于 1986 和 1989 年在重庆大学获得学士和硕士学位,现为军事医学科学院卫生装备研究所研究员,主要研究方向为测控技术与装备.
E-mail:wutaihu@vip.sina.com

陈　锋　男,1978 年 11 月生,山东菏泽人,现为军事医学科学院卫生装备研究所副研究员,主要从事医用电子技术与智能装备的研究工作.
E-mail:chenfenghj@163.com

程　智　男,1984 年 12 月出生于河南郑州市,现为军事医学科学院卫生装备研究所助理研究员,主要从事测控技术与仪器方向研究工作.
E-mail:chengzhitt@gmail.com

杜耀华　男,1978 年 7 月出生于河北唐山市.于 2000 年和 2006 年在国防科技大学获得学士和博士学位,现为军事医学科学院卫生装备研究所助理研究员,主要研究方向为智能仪器与光机电综合集成技术.
E-mail:qsyaohua@hotmail.com

暴洪涛　男,1990 年 05 月出生于陕西省西安市,2012 年毕业于西安交通大学应用物理专业,现为军事医学科学院卫生装备研究所硕士研究生,主要从事生物战剂侦检与医用电子设备研究工作.
E-mail:bht4255@163.com