基于密度-距离的 t 混合模型流式数据聚类*

赵其杰[1,2],柯震南[1],陶 靖[3],卢建霞[1]
(1. 上海大学机电工程与自动化学院 上海 200072;2. 上海市智能制造及机器人重点实验室 上海 200072;
3. 上海纳衍生物科技有限公司 上海 201108)

摘 要:传统流式数据采用人工设门法分析,效率低下且依赖于专家。近几年,很多自动流式数据聚类算法纷纷被提出,然而针对数据量不多且分布稀疏的小样本类群始终没有很好的解决办法。提出了一种基于密度-距离的 t-混合模型流式数据聚类优化方法,能够较好地解决小样本类群区分困难的问题。该方法通过密度-距离中心算法定位各类群的初始中心,作为 t-混合算法的初值对样本数据进行处理,通过最大似然估计求出各类群对应的样本数目,从而实现样本聚类。实验表明,与经典模型算法相比,基于密度-距离的 t-混合模型优化算法具有更好的稳定性和可靠性,对小样本类群以及混叠的类群具有较强的适应能力。

关键词: t-混合模型;密度-距离中心算法;流式细胞分析术;聚类算法
中图分类号: TH-773 TP-391 **文献标识码**: A **国家标准学科分类代码**: 510.40

Clustering based on density-distance and t mixture model in flow cytometry data

Zhao Qijie[1,2], Ke Zhennan[1], Tao Jing[3], Lu Jianxia[1]
(*1. School of Mechatronics Engineering and Automation, Shanghai University, Shanghai 200072, China; 2. Shanghai Key Laboratory of Intelligent Manufacturing and Robotics, Shanghai 200072, China; 3. Shanghai Nayan Biotechnology Co., Ltd, Shanghai 201108, China*)

Abstract: Traditionally, the flow cytometry data is analyzed manually, which is inefficient and depends on expert experiences. In recent years, a lot of automatic cluster algorithms have been proposed. However, the clustering performance is not satisfied for sparse data with a random distribution. Therefore, this paper presents an automatic clustering method based on density-distance center for t-mixture model algorithm in flow cytometry data, which is suitable for rare samples. The proposed method finds the center of each group by density-distance center algorithm and uses it as the initial value of t-mixture model to estimate the sample data by maximum likelihood estimation. Compared with the classical algorithm, the result shows that the t-mixture model based on density-distance center has better stability and reliability, and can better fit small or mixed samples.

Keywords: t-mixture model; density-distance center algorithm; flow cytometry; clustering algorithm

0 引 言

流式细胞分析术是采用流式细胞仪进行定量分析的技术。流式细胞仪使用流体动力学聚焦原理,将被分析的细胞或微粒排成一列,逐个快速地流过检测激光束,通过高精密的光学系统、电子学信号处理和计算机数据分析,测定细胞或颗粒引发的多角度散射光和多色荧光,可以在短时间内获得上万个细胞或微粒的大小、内部结构、DNA、蛋白质等物理及化学特征的仪器[1]。流式细胞术以其快速、准确、大批量、多参数分析等优点,是生物医疗领域中进行前沿科学研究的重要的基础性科研仪器;同时,也是重要的临床检验设备[2-5]。

传统上,流式数据的分析依靠有经验的人员将数据

收稿日期:2017-04 Received Date: 2017-04
*基金项目:上海市浦江人才计划(17PJ1432300)项目资助

投影至二维散点图中,然后采用区域设门的方式进行分析,被称为人工设门法。随着流式细胞术的不断发展,流式数据量成倍增加,数据的自动分析已经成为流式细胞技术未来发展的主要方向[6]。目前,流式数据的自动分析主要基于非监督学习算法[7-12],具体可分为基于概率的聚类方法以及基于空间信息的聚类方法。基于概率的聚类方法主要是有限混合模型,如基于贝叶斯信息准则的高斯混合模型算法[7],该算法仅对由正态或者近正态数据集组成的细胞类群有较好的处理能力。因此,Lo K 等人[8]提出 t-混合模型,并将非正态分布的数据转换为近正态分布来进行分析;Pyne S 等人[9]以及王先文等人[6]提出基于混合偏斜 t 分布模型方法,能较好地处理非对称分布的数据。基于空间信息的聚类方法是流式数据分析的另一类主要方法。如 Zare H 等人[10]采用谱聚类方法进行分析,但因需要对样本进行采样,可能造成信息信息丢失;王先文等人[11]基于 K-means 提出了一种快速流式数据分析方法;董明利等人[12]采用核熵成分分析(kernel entropy component analysis, KECA)算法选取流式数据中贡献度最高的两个参数作为分析数据,采用基于余弦相似度的 K-means 算法进行聚类。

虽然以上的流式数据聚类方法很多,但是对于样本量小且分布稀疏的类群并没有很好的解决办法。例如,人外周血的白细胞分类分析中,通常单核细胞占白细胞总量的 2% ~10%,嗜酸性粒细胞占白细胞总量的 1% ~6%,而淋巴细胞约占 40% ,粒细胞约占 50%。在这样的多类群聚类分析中,大样本类群与小样本类群的数量相差悬殊且相互靠近,难点是小样本类群的定位和区分。本文提出了一种基于密度-距离的 t-混合模型的流式数据聚类优化方法,将密度-距离中心算法应用到流式数据的初始聚类中心的定位上,从而能很好地解决小样本类群的区分问题。

1　基于密度-距离的 t-混合模型

1.1　密度-距离中心算法

密度-距离中心算法是基于局部密度和距离来寻找类群中心的,因而对小样本类群数据具有很好的鲁棒性与适应性。基于 Rodriguez A 等人[13]提出通过局部密度和距离寻找类群中心的思想,设计了筛选规则对密度-距离中心算法进行改进。对于待聚类的数据集 $S=\{x_1, x_2, \cdots, x_n\}$,可以定义局部密度 ρ_i 以及距离 δ_i 两个量($i\in[1,n]$)。局部密度反映了在一定区间内数据的密度,其定义如下[13]:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

其中,函数 $\chi(x)$ 如下:

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geqslant 0 \end{cases} \tag{2}$$

式中:参数 d_{ij} 表示的是 x_i 到 x_j 的距离,本文采用的是欧氏距离,下文中距离均指欧氏距离。参数 $d_c>0$ 为截断距离,根据实际样本数据预先设定,实验中均取 $d_c=5$。由式(1)可知,局部密度 ρ_i 表示的是数据集中与 x_i(排除自身)的距离小于 d_c 的数据点的个数。由于局部密度的特点,能够较好地突出小样本数据类群。对某一点的距离 δ_i 的定义是计算它到所有比其局部密度大的点的距离,取其中的最小值,具体公式如下[13]:

$$\delta_i = \min_{j:\rho_j>\rho_i}(d_{ij}) \tag{3}$$

如果这个点已经是局部密度最大的点,那么 δ_i 赋值为它到所有点的距离的最大值[13]。

$$\delta_i = \max_j(d_{ij}) \tag{4}$$

根据式(1)和(3),每一个点 x_i 都可以得到一个局部密度 ρ_i 和一个距离值 δ_i。

对于某种确定分析的流式数据,同一类实验样本的待分类类群数目是先验确定且相同的。因此,本文将类群数目设为定值 k。由于流式数据中位于边缘的噪音点的数据密度很小,因此设定一个局部密度的阈值 ρ_0 来过滤掉位于边缘局部密度很小但距离很大的噪音粒子。然后,把数据集按照距离 δ_i 从大到小的顺序进行排列。然后根据类群数目 k,从大到小依次取数据集中 k 个数据点作为待聚类的类群中心。设类群中心为 $x_{c_j}(j\in[1,k])$,c_j 表示类群中心点的标号(即为依次选取的 δ_i 的索引 i),D 表示已经选取的类群中心点的标号的集合,则其具体公式如下:

$$c_j = \underset{i:\rho_i>\rho_0,\ i\notin D}{\operatorname{argmax}}(\delta_i) \tag{4}$$

局部密度反映的是局部区域的数据密度,理想情况下一个类群内只有一个最大局部密度。因此,若某一类群中存在多个相等的最大局部密度点,可以对某一点的局部密度加上一个很小的增量,使其成为唯一的最大的局部密度点。为防止选取的类群中心实际上位于同一类群,对其进行筛选,筛选原则如下。

1)若两类群中心的距离值比较接近,且其空间欧式距离很小,则视为同一类群。

2)若两类群中心的局部密度近似相等,且两类群中心的欧氏距离很小,那么也视为同一类群。局部密度几乎相等,且欧氏距离小,说明是同一群粒子散开成两部分。

经过筛选后,即可获得较为准确的类群中心,作为 t-混合模型聚类算法的初始中心值,也就是式(6)中各个 t-分布分量密度函数的位置参数 μ_j。

1.2　t-混合模型

1.2.1　混合模型

设 $\boldsymbol{X}$ 为 p 维随机向量,且 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 为随机向量 $\boldsymbol{X}$

的 n 个 p 维随机样本观测值,且相互独立,则由 $\boldsymbol{X}$ 产生的由 k 个分量组成的多元混合模型概率密度函数[14-17]定义为:

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{i=1}^{k} \pi_i f(\boldsymbol{x};\boldsymbol{\theta}_i) \tag{5}$$

式中:k 为混合模型的分量数,$\boldsymbol{\Theta} = (\pi_1,\cdots,\pi_{k-1},\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_k)$ 为未知参数矩阵。$f(\boldsymbol{x};\boldsymbol{\theta}_i)$ 代表了第 i 个分量的概率密度函数,θ_i 为其未知参数向量;π_i 为混合比,表示第 i 个分量密度在混合模型中的比例,其满足 $\sum_{i=1}^{k} \pi_i = 1$,$\pi_i \geqslant 0$。

1.2.2 t-混合模型

若式(5)中 $f(\boldsymbol{x};\boldsymbol{\theta}_i)$ 为 t-分布,则 $f(\boldsymbol{x};\boldsymbol{\Theta})$ 为 t-混合模型。P 维 t-分布的概率密度函数[18-19]的形式为:

$$f(\boldsymbol{x};\theta_i) = t_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\upsilon) = \frac{\Gamma(\upsilon+p)}{(\pi\upsilon)^{p/2}|\boldsymbol{\Sigma}|\Gamma(\upsilon/2)}\left(1+\frac{\delta(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\upsilon}\right)^{-(\upsilon+p)/2} \tag{6}$$

式中:$\boldsymbol{\mu}$ 为位置参数,$\boldsymbol{\Sigma}$ 为正定矩阵,υ 为自由度,$\delta(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu})$,是 $\boldsymbol{x}$ 与 $\boldsymbol{\mu}$ 间的马氏距离的平方,$\Gamma(x)$ 为 Gamma 函数,定义为 $\Gamma(x) = \int_0^{+\infty} \mathrm{e}^{-t}t^{x-1}\mathrm{d}t$,$x>0$。对于 t 混合模型,每个分量密度函数都为 P 维 t-分布密度函数,其混合模型式为:

$$f(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{i=1}^{k} \pi_i t_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\upsilon) \tag{7}$$

对于流式数据,若其可分为 k 个类群,则 t-混合模型假定它是由 k 个 t-分布组成的。最后的聚类结果也就是求出对应 k 个 t-分布的 k 个流式细胞群。通过对流式数据样本建立极大似然估计,采用 EM(expectation maximization)算法可获得极大似然估计的混合参数。$\boldsymbol{X}_i$ 为流式数据中的某一个 p 维样本值,$\boldsymbol{X}_i = (x_{i1}, x_{i2},\cdots, x_{ip})^{\mathrm{T}}$。引入 $\boldsymbol{X}_i$ 分量的标记向量 $\boldsymbol{Z}_i = (z_{i1}, z_{i2},\cdots, z_{ik})^{\mathrm{T}}$,且满足:$\boldsymbol{X}_i$ 属于第 j 个 t-分布时,$z_{ij}=1$,否则 $z_{ij}=0$。即 $\boldsymbol{Z}_i$ 表示该样本值 $\boldsymbol{X}_i$ 属于哪一个 t-分布。此时,完全数据向量集为 $\boldsymbol{X}_C = (\boldsymbol{X}^{\mathrm{T}},\boldsymbol{Z}_1^{\mathrm{T}},\boldsymbol{Z}_2^{\mathrm{T}},\cdots,\boldsymbol{Z}_n^{\mathrm{T}})^{\mathrm{T}}$。其中 $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}},\cdots, \boldsymbol{X}_n^{\mathrm{T}})^{\mathrm{T}}$。其相应的对数似然函数可以写为:

$$\ln(L(\boldsymbol{\Theta}|X_c)) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}(\ln(f(\boldsymbol{x};\theta_j)) + \ln(\pi_j)) \tag{8}$$

1.2.3 EM 算法估计

对于 t-混合模型,采用 EM 算法[8,19]进行参数估计的过程如下:

1)E 阶段:设 $\boldsymbol{\Theta}^{(t)}$ 为第 t 次迭代的估计值,则在给定条件 $\boldsymbol{\Theta}^{(t)}$ 下的对数似然函数的条件期望如式(9)。

$$Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)})) = E(\ln(L_c(\boldsymbol{\Theta}|\boldsymbol{X}_c));\boldsymbol{\Theta}^{(t)}) \tag{9}$$

2)M 阶段:由式(8)求 $\boldsymbol{\Theta}^{(t+1)}$ 使 $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t+1)})$ 最大。

$$\boldsymbol{\Theta}^{(t+1)} = \mathrm{argmax}(Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(t)})) \tag{10}$$

3)由式(9)和(10)循环迭代直至参数收敛,得到参数 $\boldsymbol{\Theta}$ 的估计值。

由 EM 算法求得的相应参数的迭代式为:

$$\boldsymbol{\pi}_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(t+1)} \tag{11}$$

$$\boldsymbol{\omega}_{ij}^{(t+1)} = \frac{\upsilon_j^{(t)}+p}{\upsilon_j^{(t)}+\delta(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{12}$$

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t+1)}\boldsymbol{\omega}_{ij}^{(t+1)}x_i}{\sum_{i=1}^{n} z_{ij}^{(t)}\boldsymbol{\omega}_{ij}^{(t)}} \tag{13}$$

$$\boldsymbol{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} z_{ij}^{(t+1)}\boldsymbol{\omega}_{ij}^{(t+1)}(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(t+1)})(\boldsymbol{x}_i-\boldsymbol{\mu}_j^{(t+1)})^{\mathrm{T}}}{\sum_{i=1}^{n} z_{ij}^{(t)}} \tag{14}$$

自由度 $\upsilon_j^{(t+1)}$ 是非线性方程:

$$\boldsymbol{\Psi}\left(\frac{\upsilon_j^{(t)}+p}{2}\right) - \boldsymbol{\Psi}\left(\frac{\upsilon_j^{(t+1)}}{2}\right) + \ln\left(\frac{\upsilon_j^{(t+1)}}{2}\right) + 1 + \frac{\sum_{i=1}^{n} z_{ij}^{(t)}(\ln(\boldsymbol{\omega}_{ij}^{(t)}) - \boldsymbol{\omega}_{ij}^{(t)})}{\sum_{i=1}^{n} z_{ij}^{(t)}} - \ln\left(\frac{\upsilon_j^{(t)}+p}{2}\right) = 0 \tag{15}$$

的解,其中 $\boldsymbol{\Psi}(x) = \frac{\mathrm{d}\ln(\Gamma(x))}{\mathrm{d}x}$。

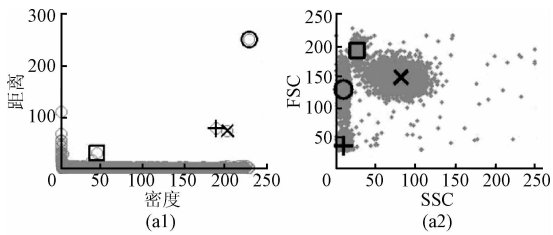
2 实验与分析

在流式数据的分析中,应用单克隆抗体-荧光标记获得的数据由于抗体的特异性结合的特征,二维散点图分类明显,往往较易区分;基于前向散射光和侧向散射光的数据分析在流式细胞仪和全自动血液分析仪中被广泛采用,例如白细胞的分类分析,因为相比较荧光检测,散射光的检测是成本更经济的方案。但是,基于前向和侧向散射光的流式数据,因为类群分离度受样本处理、噪音、仪器灵敏度的影响,数据类群之间往往界限模糊,因此分析更加困难。为验证本文提出的方法对于实际样本数据的自动分析能力,在上海纳衍生物科技有限公司研发的多色荧光流式细胞分析仪上,对从某医院临床获取的成人外周血样本进行分析。实验抽取前向散射光通道(forward scatter, FSC)和侧向散射光通道(side scatter, SSC)的数据建立二维散点图(横轴为侧向散射光,纵轴为前向散射光),用于算法分析实验。

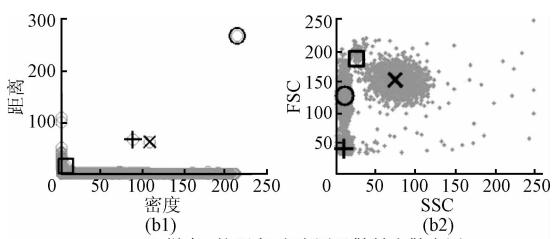
实验 1 分析了密度-距离中心算法的稳定性与鲁棒性;实验 2 对比分析了本文算法以及经典算法 flowPeaks、flowClust 针对小样本类群的聚类能力;实验 3 通过与人工设门法进行对比,定量分析了本文算法的准确性。本文提出的算法流程如下:1)根据密度-距离中心算法初始化各个类群的中心;2)将各类群的中心作为 t-混合模型

的初值进行聚类。

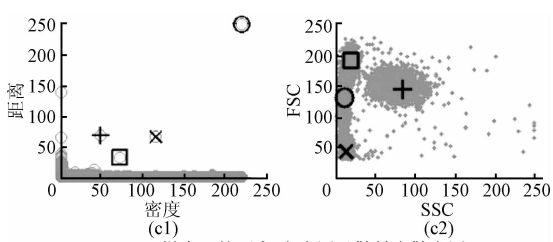
实验 1 通过对大量样本重复实验分析了密度-距离中心算法初始化类群中心的效果。实验样本来自医院成年人外周血样本(包含健康人及病患血样),其中包含淋巴细胞、单核细胞、粒细胞和红细胞碎片等几大类群。本文特意选取其中最具有代表性的样本进行分析。图 1(a1)~(e1)所示为细胞群的距离-密度分布,根据分布图由距离从大到小选取类群中心,分别用“○△”、“+”、“△”和“□”表示;图 1(a2)~(e2)所示为原图中对应类群中心的位置。图 1 中,样本 I 单核细胞群约占 5%,各类群区分明显,左上方为淋巴细胞群,左下方为红细胞碎片,中上方为单核细胞群,右方为粒细胞群;样本(II)中单核细胞群样本量很少,约占 2%,为病患或极端情况;样本 III 中单核细胞群(样本量较多)与淋巴细胞群靠得很近;样本 IV、V 中单核细胞群不仅样本量少(约占 2%),而且与淋巴细胞群靠得很近,部分混叠。这些对分类算法提出了很大的挑战。



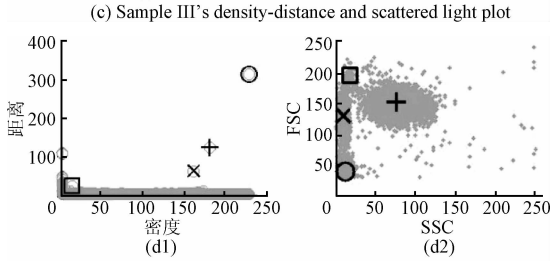
(a) 样本I的距离-密度图及散射光散点图
(a) Sample I's density-distance and scattered light plot



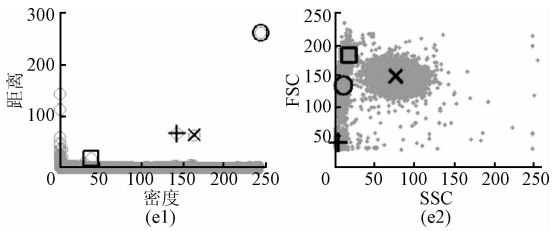
(b) 样本II的距离-密度图及散射光散点图
(b) Sample II's density-distance and scattered light plot



(c) 样本III的距离-密度图及散射光散点图
(c) Sample III's density-distance and scattered light plot



(d) 样本IV的距离-密度图及散射光散点图
(d) Sample IV's density-distance and scattered light plot



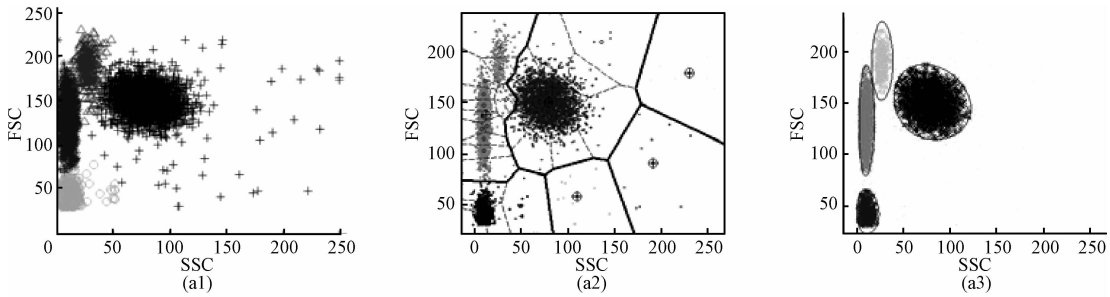
(e) 样本V的距离-密度图及散射光散点图
(e) Sample V's density-distance and scattered light plot

图 1　初始化类群中心

Fig. 1　Initializing cluster centers

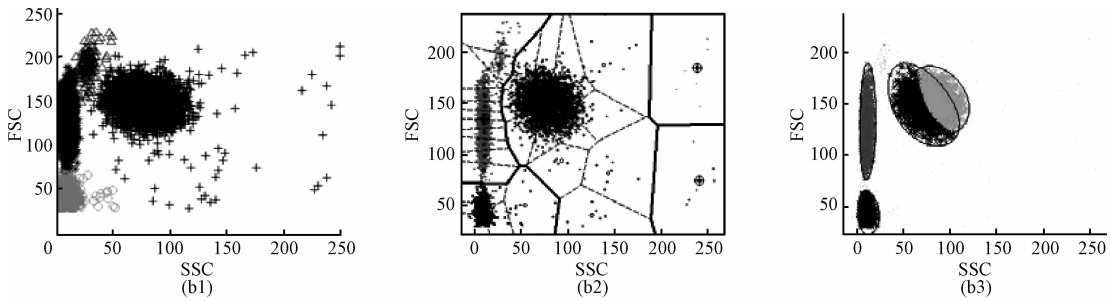
在上述实验样本中,既有小样本类群,又有互相靠近的类群,这些对分类算法提出了很大的挑战。由图 1 可知,密度-距离中心算法对区分小样本类群以及相互靠近的类群均有出色的表现。多次实验,对各种恶劣分布情况,结果均很稳定。前人大多采用改进的 K-means 算法初始化聚类中心,虽然采用了很多条件进行改进与限制,但是 K-means 算法本身求得的是局部最优解,因此对于随机的初值依然有可能陷入局部最优,造成误分。然而,采用密度-距离中心算法求出的类群中心点非常稳定,作为混合模型的初值,能有效地避免陷入局部最优。可见,基于密度-距离中心算法的初始化聚类中心方法更加稳定可靠。

为验证对小样本类群的分析能力,实验 2 对样本 a~j 分别用本文算法、flowPeaks 以及 flowClust 做对比,图 2(a1)~(i1)所示为本文算法的处理结果,类群数据分别用“*”、“o”、“+”和“Δ”表示。图 2(a2)~(i2)所示为 flowPeaks 算法的处理结果,各类群由粗实线隔开。图 2(a3)~(i3)所示为 flowClust 的处理结果,划分后的类群由椭圆形圈出,间隔可由颜色深浅区分。由图 2 可知,flowPeaks 算法仅对样本 f 聚类成功,其对小样本类群以及相互靠近类群的聚类效果差。这是因为 flowPeaks 首先采用基于距离的 kmeans++产生初始点,最后会根据局部峰的高度来合并各类群,小样本类群密度低,局部峰低,会被合并为最近的局部峰高的类群。flowClust 也对小样本类群的聚类效果不好。这是因为 flowClust 采用 Box-Cox 变换(将非正态数据转换成近正态数据),对于实验样本 b~e、g 和 h 中少于 3% 的类群,经变换后密度更低,从而被误分为噪音。由于待分类样本少一个类群,flowClust 聚类结果难免出错。本文算法采用先定位类群中心再进行--混合模型聚类的思路,有效地避免了这种小样本粒子群误分。由此可见,本文算法的确对小样本流式数据有比较强的适应能力。根据血细胞的流式数据的分布特点以及算法效率,本文选用了 t-混合模型,t-混合模型抗干扰能力强,能较少对周围强噪音点的误分。



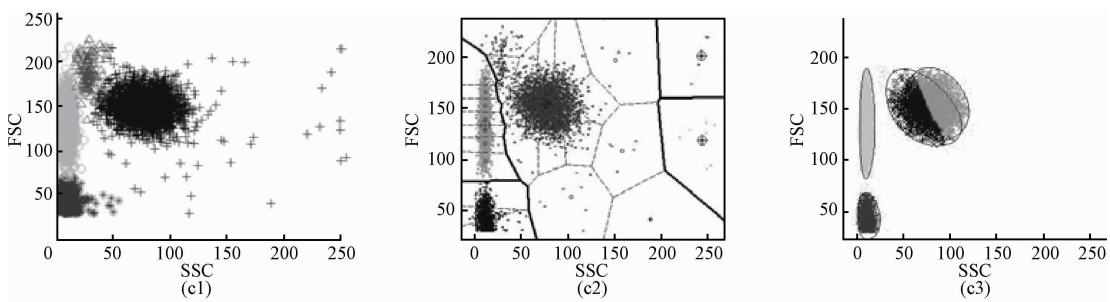
(a) 本文算法、flowPeaks、flowClust对样本a的聚类结果

(a) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample a



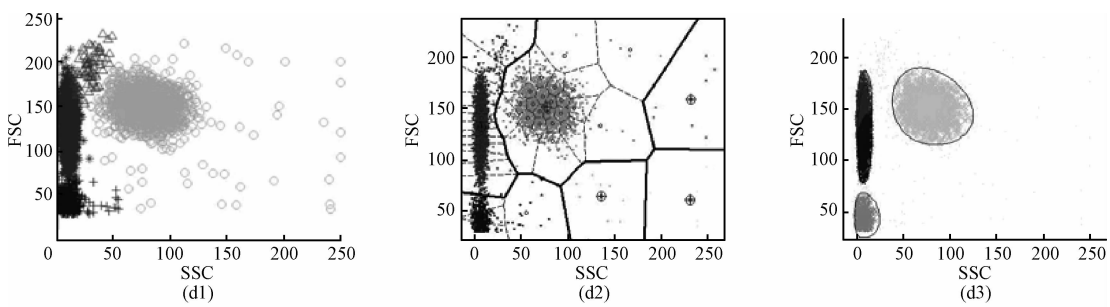
(b) 本文算法、flowPeaks、flowClust对样本b的聚类结果

(b) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample b



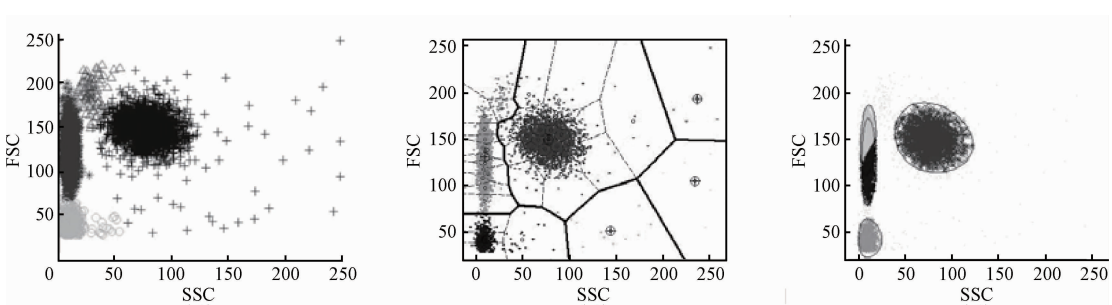
(c) 本文算法、flowPeaks、flowClust对样本c的聚类结果

(c) Clustering results of the propsoed algorithm, flowPeaks and flowClust for sample c



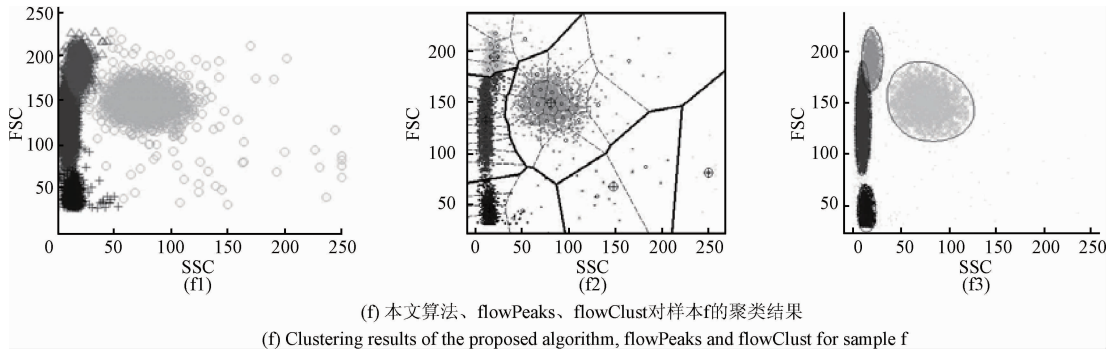
(d) 本文算法、flowPeaks、flowClust对样本d的聚类结果

(d) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample d



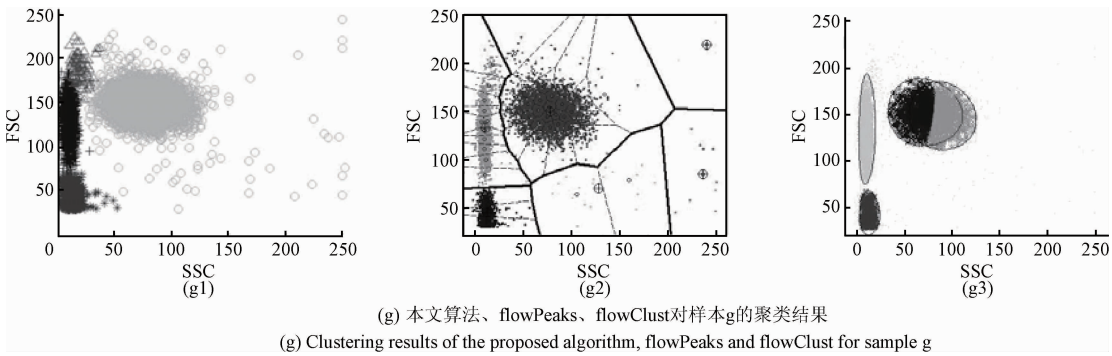
(e) 本文算法、flowPeaks、flowClust对样本e的聚类结果

(e) Clustering results of proposed algorithm, flowPeaks and flowClust for sample e



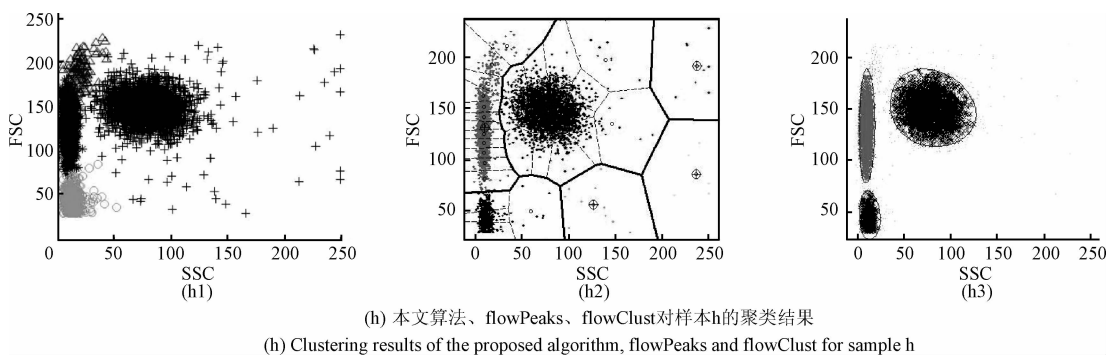
(f) 本文算法、flowPeaks、flowClust对样本f的聚类结果

(f) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample f



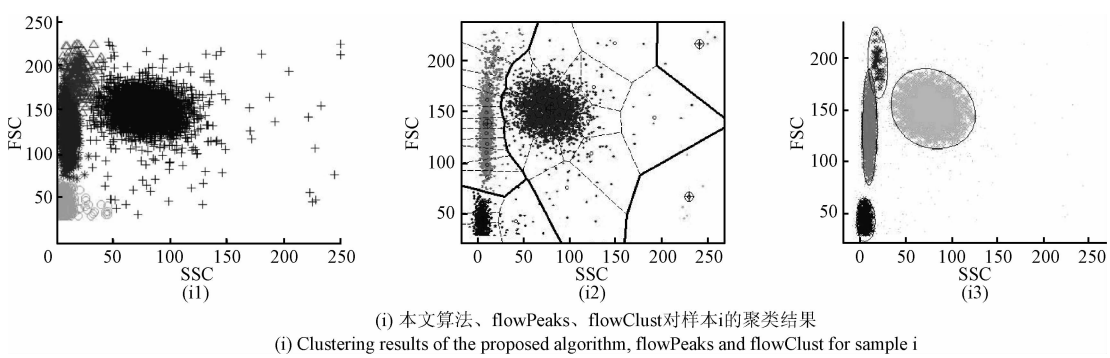
(g) 本文算法、flowPeaks、flowClust对样本g的聚类结果

(g) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample g



(h) 本文算法、flowPeaks、flowClust对样本h的聚类结果

(h) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample h



(i) 本文算法、flowPeaks、flowClust对样本i的聚类结果

(i) Clustering results of the proposed algorithm, flowPeaks and flowClust for sample i

图 2　三种算法分析对比图

Fig. 2　Comparison and analysis results of three algorithms

实验 3 以人工设门法的分类结果为标准,分析了本文算法对样本 a ~ j 的聚类效果。聚类后的样本分为 4 群,分别为红细胞碎片、淋巴细胞、单核细胞以及粒细胞。表 1 所示为用密度-距离中心算法优化的 t-混合模型算法聚类后的粒子类群相对于人工设门法的误差百分比(误差百分比 = 本文算法的粒子群占粒子总数百分比-人工设门法的对应粒子群占粒子总数百分比),最后一行为每一列对应的平均值。对于粒子数较少的单核细胞,本文算法的误差平均值仅为 0. 201%,可见本文提出的基于密度-距离中心算法优化的 t-分布混合模型对小样本类

群分类准确率高,能有效地解决小样本流式数据的聚类问题。对于大约包含 5 000 个细胞的样本流式数据,在 MATLAB 平台下,本文算法平均运行时间为 43.42 s。

表 1 分类结果表

Table 1 Classification results

样本	红细胞碎片	淋巴细胞	单核细胞	粒细胞	时间/s
a	0.2	0.23	0.33	0.27	34.57
b	0.07	0.18	0.05	0.04	41.16
c	0.17	0.28	0.04	0.02	41.75
d	0.21	0.1	0.33	0.09	35.92
e	0.11	0.06	0.19	0.05	35.96
f	0.31	0.26	0.24	0.21	45.78
g	0.04	0.08	0.27	0.02	57.27
h	0.2	0.2	0.31	0.01	42.39
i	0.05	0.35	0.14	0.08	50.91
j	0.02	0.26	0.11	0.24	48.56
平均值	0.138	0.2	0.201	0.103	43.42

3 结 论

本文提出了一种基于密度-距离的 *t*-混合模型优化算法,能在已知类群数目下,实现对小样本类群以及混叠的流式数据的准确分析。本文根据先定位再聚类的思想,尝试了一种新的类群初始中心定位方法,能准确地确定各类群中心的大致位置。以往研究中,常用 K-means 来初始化类群中心,通过尽量使各初始聚类中心的相互距离尽可能地远来避免局部最优,但依然可能找到边缘噪音点。本文通过密度-距离中心算法先确定类群中心的方法具有很好的稳定性和可靠性,避免了这种不足;实验 2 与经典流式算法 flowPeak、flowClust 进行对比,表明了基于密度-距离的 *t*-混合模型优化算法对小样本类群以及混叠的类群数据具有较强的适应能力;实验 3 的误差分析也验证了本文算法的确对小样本类群具有较高的聚类能力,误差小。

参考文献

[1] 裴智果, 王策, 陈忠祥,等. 用于流式细胞仪的数据采集系统设计与实现[J]. 电子测量技术, 2015,38(7): 84-88.
PEI ZH G, WANG C, CHEN ZH X, et al. Design and implementation of data acquisition system for flow cytometry [J]. Electronic Measurement Technology, 2015,38(7): 84-88.

[2] 吴云良, 裴智果, 陈忠祥, 等. 以 FPGA 为核心的流式细胞仪控制系统设计[J]. 电子测量技术, 2015, 38(7): 58-61.
WU Y L, PEI ZH G, CHEN ZH X, et al. Design of flow cytometry control system based on FPGA[J]. Electronic Measurement Technology, 2015,38 (7): 58-61.

[3] 张文昌, 祝连庆, 娄小平,等. 基于灰色预测恢复算法的流式细胞仪多参数提取[J]. 仪器仪表学报, 2015, 36(7):1660-1665.
ZHANG W CH, ZHU L Q, LOU X P. Multi-parameter extraction of flow cytometer based on grey prediction recovery algorithm [J]. Chinese Journal of Scientific Instrument, 2015, 36(7):1660-1665.

[4] PEDREIRA C E, COSTA E S, LECREVISSE Q, et al. Overview of clinical flow cytometry data analysis: Recent advances and future challenges [J]. Trends in Biotechnology, 2013, 31(7):415-425.

[5] 程振, 杨斌, 徐友春,等. 用于流式细胞仪的超声聚焦系统的仿真与设计[J]. 仪器仪表学报, 2017, 38(6): 1547-1553.
CHEN ZH, YANG B, XU Y CH, et al. Simulation and design of an acoustic focusing system for flow cytometer[J]. Chinese Journal of Scientific Instrument, 2017, 38(6): 1547-1553.

[6] 王先文, 陈锋, 程智,等. 基于偏斜 *t* 混合模型的流式数据自动聚类方法研究[J]. 电子学报, 2014, 42(12):2527-2535.
WANG X W, CHEN F, CHENG ZH, et al. Auto clustering method study of flow cytometry data based on skew *t*-mixture models [J]. Chinese Journal of Electronics, 2014, 42(12): 2527-2535.

[7] NIMA A, GREG F, HOLGER H, et al. Critical assessment of automated flow cytometry data analysis techniques [J]. Nature Methods, 2013, 10 (3): 228-238.

[8] LO K, BRINKMAN R R, GOTTARDO R. Automated gating of flow cytometry data via robust model-based clustering[J]. Cytometry Part A, 2008, 73A (4): 321-332.

[9] PYNE S, HU X, WANG K, et al. Automated high-dimensional flow cytometric data analysis [J]. Proceedings of the National Academy of Sciences, 2010, 106(21):8519-8524.

[10] ZARE H, SHOOSHTARI P, GUPTA A, et al. Data reduction for spectral clustering to analyze high throughput flow cytometry data[J]. BMC Bioinformatics, 2010, 11(1):1-16.

[11] 王先文, 王懿男, 暴洪涛,等. 一种快速自动分析流式数据方法研究[J]. 军事医学, 2015, 39(10): 736-741.

WANG X W, WANG Y N, BAO H T, et al. A rapid and automatic flow data analysis method[J]. Military Medical Sciences, 2015, 39(10): 736-741.

[12] 董明利, 马闪闪, 张帆, 等. 基于核熵成分分析的流式数据自动分群方法[J]. 仪器仪表学报, 2017, 38(1): 206-211.
DONG M L, MA SH SH, ZHANG F, et al. Automatic clustering method for streaming data based on kernel entropy component analysis[J]. Chinese Journal of Scientific Instrument, 2017, 38(1): 206-211.

[13] RODRIGUEZ A, LAIO A. Machine learning: Clustering by fast search and find of density-distance centers[J]. Science, 2014, 344(6191):1492-1496.

[14] 刘恒, 吴迪, 苏家仪,等. 运用高斯混合模型识别动物声音情绪[J]. 国外电子测量技术, 2016, 35(11): 82-87.
LIU H, WU D, SU J Y, et al. Recognition of animal sound's emotion based on Gaussian mixture model[J]. Foreign Electronic Measurement Technology, 2016, 35(11): 82-87.

[15] 李菊, 李克清, 苏勇刚. Markov 随机游走和高斯混合模型相结合的运动目标检测算法[J]. 电子测量与仪器学报, 2014, 28(5):533-537.
LI J, LI K Q, SU Y G. Moving target detection algorithm combined with Markov random walk and Gauss mixed model[J]. Journal of Electronic Measurement and Instrument, 2014, 28(5):533-537.

[16] GAO C, ZHU Y, SHEN X, et al. Estimation of multiple networks in Gaussian mixture models[J]. Electronic Journal of Statistics, 2016, 10(1):1133-1154.

[17] MALSINER-WALLI G, FRüHWIRTH-SCHNATTER S, GRüN B. Model- based clustering based on sparse finite Gaussian mixtures[J]. Statistics and Computing, 2016, 26(1-2): 303-324.

[18] CASTRO L M, COSTA D R, PRATES M O, et al. Likelihood-based inference for Tobit confirmatory factor analysis using the multivariate Student-t distribution[J]. Statistics and Computing, 2015, 25(6): 1163-1183.

[19] GHOSH A K, CHAKRABORTY A. Use of EM algorithm for data reduction under sparsity assumption[J]. Computational Statistics, 2017, 32(2): 387-407.

作者简介

赵其杰(通讯作者),2005 年于上海大学获得博士学位,现为上海大学副教授,主要研究方向为传感检测与控制、机器视觉、人机交互与智能信息处理。
E-mail:zqj@ shu. edu. cn

Zhao Qijie (Corresponding author) received his Ph. D. degree from Shanghai University in 2005. Now he is an associate professor in Shanghai University. His main research interests include Measurements and control with sensors, machine vision, Human-robot interaction, and intelligent information processing.

柯震南,2015 年于上海大学获得学士学位,现在上海大学硕士研究生,主要研究方向为机器视觉。
E-mail: kezhennan@126. com

Ke Zhennan received his B. Sc. degree from Shanghai University in 2015. Now he is a master candidate in Shanghai University. His main research interest is Computer Vision.