面向结构稳定性的分裂-合并聚类算法

雷小锋[1] 何 涛[1] 李奎儒[1] 谢昆青[2] 丁世飞[1]
(中国矿业大学计算机科学与技术学院 徐州 221116)[1]
(北京大学信息科学技术学院视觉与听觉国家重点实验室 北京 100871)[2]

摘 要 聚类是在假设数据具有某种群聚结构的前提下根据观察到的无标记样本发现数据的最优划分。现有的聚类算法通常简单地导出假设结构和给定先验下最优或较优的聚类结果,体现为算法对样本分布拟合度的迭代最优化,即算法有效性。实际上,聚类的有效性取决于结构有效性、算法有效性和先验有效性3个方面的因素。基于这种考虑,提出了一种变体混合模型的聚类结构假设,以及判定聚类结构的稳定性的度量和方法,在算法有效的前提下通过单簇的分裂与合并来改进聚类结构的稳定性,并得到最终聚类结果,设计并实现了SMClus聚类算法,通过对模拟数据和真实数据的聚类实验,例证了方法的有效性。

关键词 聚类算法,变体混合模型,结构稳定性,分裂-合并
中图法分类号 TP181 **文献标识码** A

Split-Merge Based Clustering Algorithm Oriented to Structure Stability of Clusters

LEI Xiao-feng[1] HE Tao[1] LI Kui-ru[1] XIE Kun-qing[2] DING Shi-fei[1]
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)[1]
(Department of Intelligence Science/National Laboratory on Machine Perception, Peking University, Beijing 100871, China)[2]

Abstract Clustering is to find the best partition of unlabeled observations under a certain group stucture hypothesis. Given the group stucture hypothesis, the most clustering algorithms is to to iteratively optimize of fittness of data distribution (called algorithm validity). In fact, the clustering validity is determined by three factors: hypothesis, algorithm and apriori validity. Therefore, a variation of gaussian mixture model was proposed in this paper, then the measurement and estimation method of cluster structure stability were defined. Based on them, the SMClus algorithm was designed to achieve the stable clustering structure by means of split-merge operations. The experiment shows SMClus' performance in clustering quality.

Keywords Clustering, Variation of mixture model, Structure robustness, Split-Merge

聚类是一种无监督的分类,在文献[1,12]中对现有聚类方法进行了很好的综述。能否从大量的未标记样本中获得一些有用的信息,完全取决于我们是否接受向数据强加一些假设。实际中,不同的聚类方法都会隐含地为数据强加一个结构,尽管该结构有可能不存在。那么,在类标记未知的情况下,如何才能知道算法产生的聚集结构是否适合于数据集,而非所选择的特定聚类方法强加给数据的,即聚类结果的有效性问题。一个简单的方法是在低维投影空间中观察数据,这种方法比较适合于维度较低的情况。但是,对于高维数据,如文档的关键词TF/IDF向量,此时观察方法的有效性值得怀疑。

事实上,影响聚类结果有效性的因素有3个:其一是聚类方法强加给数据的结构模型是否有效,称为假设结构有效性;其二是在假设结构下特定的聚类方法能否导出与数据匹配度很高的结果,即算法有效性;其三是先验知识的有效性,如用户给定的类别数是否正确。因此,聚类结果的有效性$V(R)$可以定义为:

$$V(R)=V(P)\times V(S|P)\times V(A|S)$$

式中,$V(P)$为先验知识的有效性,$V(S|P)$为给定先验知识条件下假设结构的有效性,$V(A|S)$为算法的有效性。在特定假设结构下,通常的聚类算法都能够很好地优化数据拟合度,保证算法的有效性。聚类算法一般假定用户可以保证先验知识的有效性,但是大部分情况下这种假设并不合理,或许用户对数据有一定程度的认识,通常并不准确。而结构有效性则首先要求算法所假设的结构一定是蕴含了实际数据的聚集结构,否则根本不可能获得有效的聚类结果,此外结构有效性一定程度地依赖于先验知识,如K-Means算法的聚类结构就依赖于给定的类别数目。

到稿日期:2009-12-08 返修日期:2010-03-01 本文受863国家高技术研究发展计划(2006AA12Z217)和中国矿业大学科技基金(No. OD080313)资助。

雷小锋(1975-),男,博士后,主要研究方向为时空数据库与时空数据挖掘、机器学习等,E-mail:leiyunhui@gmail.com;何 涛(1987-),男,硕士生,主要研究方向为自然语言处理数据挖掘等;李奎儒(1986-),男,硕士生,主要研究方向为图像处理等;谢昆青(1941-),男,博士生导师,主要研究方向为智能信息处理;丁世飞(1963-),男,博士后,主要研究方向为人工智能与模式识别。

因此，忽略先验知识和假设结构的有效性，孤立地考虑聚类算法的有效性通常会产生无效的聚类结果。这恰恰是大部分现有聚类方法的问题所在，即缺乏对聚类结果的自省能力，只是简单地导出假设结构下最优的聚类结果，并交付用户。实际上，聚类方法是一个不断迭代优化的过程，不仅仅是对训练数据拟合度的优化，更重要的是对假设结构和先验知识的迭代优化。基于上述的考虑，本文引入了一种变体混合模型的聚类结构假设，以及判定聚类结构的稳定性的度量和方法，并基于分裂-合并策略提出了一种面向结构稳定性的 SMClus 聚类算法(Split-Merge based CLUStering algorithm)，通过对模拟数据和真实文档数据的聚类实验，例证了本文方法的有效性。

本文第 1 节对现有的一些聚类算法进行了综述；第 2 节给出了聚类结构假设模型及其稳定性判定和度量方法；第 3 节详细说明 SMClus 算法；第 4 节是性能分析和实验测试；最后是结论和下一步工作说明。

1 相关工作

划分聚类算法通过迭代重定位策略优化特定的目标函数，尝试确定数据集的一个划分。最常用的目标函数是误差平方和准则，如 K-Means 算法[2]。K-Means 算法对类球形且大小差别不大的类簇有很好的表现，但不能发现形状任意和大小差别很大的类簇，且聚类结果易受噪声数据影响。此外，K-Means 算法仅保证快速收敛到局部最优结果，导致聚类结果对初始代表点的选择非常敏感。

层次聚类算法以自顶向下(分裂)或自底向上(凝聚)的方式将数据对象划分成一个层次树结构，即类簇树。算法的聚类效果很大程度上依赖于度量类簇之间相异度的距离函数，此外一般层次聚类算法的伸缩性不强，其时间复杂度通常为 $O(n^2)$。BIRCH[3]和 CURE[4]算法试图提高层次聚类结果的质量，解决其算法伸缩性问题。BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)是一种高效的分裂式层次聚类算法，具有 $O(n)$的计算复杂度，在有限内存下可以很好地工作。但是，算法使用的相异度度量导致 BIRCH 只能发现球形类簇。CURE(Clustering Using REpresentatives)使用固定数量的代表点来定义类簇，可以发现复杂形状和不同大小的类簇，对噪声具有很好的免疫能力。CURE 利用随机采样和划分技术处理大型的数据集，测试结果表明它较 BIRCH 性能更优。然而，CURE 算法的收缩方式隐含地依赖于球形类簇假设，故在处理特殊形状的类簇时比较困难。Chameleon[5]算法则通过图划分算法将样本对象聚类为大量相对较小的子簇，然后进行层次聚类。

基于密度的聚类算法中类簇被定义为连通的稠密子区域(component)，因此，算法能够发现任意形状的类簇，并对异常点和噪声有自然的免疫能力。DBSCAN (Density Based Spatial Clustering of Applications with Noise)是典型的面向低维空间数据聚类的基于密度的算法[6]，其关键概念是由对象最近邻域的局部分布度量的密度及其连通性。DBSCAN 算法本质上只是提供了一个根据密度阈值参数进行聚类结果搜索的过程，聚类结果在用户指定密度阈值那一刻已经唯一地确定，算法本身并不对聚类的结果负责。OPTICS[7]算法针对 DBSCAN 算法的缺陷进行了改进。在聚类计算之前，OPTICS 算法首先将基于密度计算类簇所需的信息记录下来，这些信息反映了基于密度的聚类结构。基于这些信息，用户可以比较容易地确定合适的密度参数阈值。

2 聚类结构假设及其稳定性评估

根据结构有效性的含义，首先假设的聚类结构一定要涵盖实际数据的分布结构，即假设的聚类结构要有足够的表达能力，K-Means 算法的假设就限制了其只能够发现指定数目的类球形且大小差别不大的类簇结构。这里，我们采用一种从一般混合模型导出的变体混合模型来作为聚类结构的假设。

通用混合模型假设样本独立地来自于 k 类分布组成的混合模型，每种类别的先验概率为 $P(C_j)$，然后根据概率密度 $p(x|\theta_j,C_j)$生成具体的样本，即样本来自于如下形式的概率模型：

$$p(x|\Theta)=\sum_{j=1,k}P(C_j)p_j(x|\theta_j,C_j) \quad (1)$$

式中，$\Theta=(\theta_1,\cdots,\theta_k)$是待估计的参数向量；条件概率密度 $p(x|\theta_j,C_j)$称为分量密度，表示类别 j 的概率密度形式；先验概率 $P(C_j)$又称为混合因子。

2.1 变体混合模型假设

混合模型假设的表达能力足够强大，但求解起来也非常困难。为此，K-Means 算法对其进行了简化，要求每个类别的概率密度形式为协方差矩阵相同的球形高斯分布，即 $\theta_j=(\mu_j,\Sigma)$且 $\Sigma=\sigma^2 I$，μ_j,σ^2 未知，并假设所有类别的混合因子相等，从而获得高性能的求解过程，但这种简化假设自然地限制了模型的表达能力，使得 K-Means 算法无法处理任意形状和尺寸的类簇。

这里，对原始混合模型和 K-Means 的简化假设进行了折中。首先保持原有的混合模型假设：样本独立地来自于 k 类概率分布组成的混合模型，每个类别均有相应的混合因子和分量密度，但是不再限制球形高斯分布和相等的混合因子，而是假设 k 类概率分布中的每个类别 C_j 的概率密度均可表示为 m_j 个球形高斯密度的混合模型，即变体混合模型假设：

$$\begin{aligned}p(x|\Theta)&=\sum_{j=1,k}P(C_j)p_j(x|\theta_j,C_j)\\&=\sum_{j=1,k}\sum_{i=1,m_j}P(C_j)P(C_i)p_i(x|\mu_i,\Sigma_n)\\&=\sum_{l=1,n}\alpha_l p_l(x|\mu_l,\Sigma_l)\end{aligned} \quad (2)$$

$$n=\sum_{j=1,k}m_j,\ \alpha_l=P(C_j)P(C_i)$$

式中，α_n 即为假设概率模型的混合因子。可以看出，模型最终表示为 n 个球形高斯分布的加权，每个高斯分量的协方差矩阵不要求相等，即 $\theta_j=(\mu_j,\Sigma_j)$且 $\Sigma_j=\sigma_j^2 I$。很明显，只要每个类别的高斯分量足够多，整个模型就可以用来近似任意的数据分布，可以涵盖非常复杂的数据分布结构，因此这种聚类结构的假设具有足够强的表达能力。

这里，变体混合模型中的一个高斯分量描述了一组样本点的分布，称这组样本点组成的集合为一个单簇，多个单簇组成一个类簇，最终的变体混合模型对应于 n 个单簇、k 个类簇的聚类结构。不过，这个变体混合模型求解起来依然比较困难。本文求解的基本思路是：给定一个变体混合模型所蕴含的初始聚类结果(假设 r 个类簇，每个类簇包含 1 个单簇)，然后对 r 个类簇中的每个单簇进行结构稳定性评估，若不稳定则通过分裂操作来提升其簇内稳定性，直到每个单簇都稳定为止，得到 n 个单簇，即算法的分裂阶段；此后，对两两单簇进

行簇间稳定性评估，并基于评估结果对单簇进行合并操作，得到 k 个类簇的聚类结果，即算法的合并阶段。整个算法是一种面向结构稳定性的分裂-合并聚类过程，具体过程如图 1 所示。

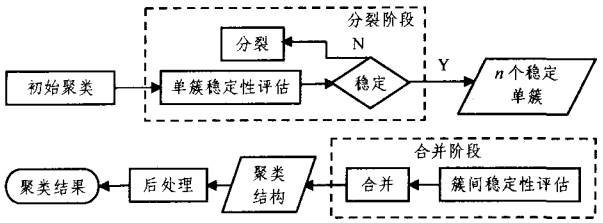

图 1　面向结构稳定性的分裂-合并聚类过程

在面向结构稳定性的聚类过程中，需要使用某种聚类算法来完成初始聚类结构的生成、单簇稳定性评估以及一个高斯分量所对应单簇的分裂等工作，最直接的想法是采用经典的 K-Means 算法。但是，面向结构稳定性的聚类过程并不奢望在一次迭代中解决问题，所以无需集成像 K-Means 这样偏置过强且影响整体性能的算法，只需要一次分裂中的结果比随机猜测稍好即可，称这种聚类算法为“弱聚类器”。这里实际上借鉴了集成机器学习(Ensemble Learning)[8,9]的思想，即多个弱分类器(性能仅比随机猜测好)可以组合为一个强分类器。同理，利用多个弱聚类器的结果为稳定性评估提供证据，进行迭代分裂和合并，最终可以达到增强聚类结构稳定性的目的。关键问题是如何建立聚类结构的稳定性评估机制。在文献[11]中我们提出 KMeanSCAN 算法，其中利用 K-Means 算法的局部最优性和初始代表点敏感性评估聚类结构的稳定性。

2.2　聚类结构的稳定性评估

聚类结构由一系列稳定的单簇组成，因此评判其结构稳定性的基本思路是：如果组成该聚类结构的单簇存在很大程度的不稳定倾向，则整个结构也就不稳定。具体到某个单簇而言，实际上有两方面的含义：

(1)单簇的簇内稳定性：即描述该单簇的高斯分量是否能够很好地表达单簇局部的数据分布；

(2)单簇的簇间稳定性：即两个单簇是否具有很强的归并于同一个类别的倾向。

假设一个变体混合模型具有 k 个类别，每个类别由 m_j 个球形高斯分量混合而成，密度参数分别为 (μ_i,σ_i)，$i=1\cdots m_j$，每个球形高斯分量描述一个单簇。那么，如何评估给定单簇的簇内稳定性呢？首先给出单簇密度的定义和一个结论。

定义 1(单簇密度)　一个单簇 s 是由一个高斯分量描述的一组样本点的集合，其单簇密度定义为：

$$den(s)=\|s\|/\sigma^d \tag{3}$$

式中，$\|s\|$ 表示集合 s 的基数，σ 表示单簇的球形高斯分量的标准差，d 为样本点的维度。对于一个稳定单簇 s，即该单簇的高斯分量可以很好地表达单簇局部的数据分布的情况，有以下结论：

结论 1　如果单簇 s 簇内稳定，且假设对 s 可以进行一次或多次聚类分裂而得到 g 个单簇 $\{s_1,\cdots,s_g\}$，若 g 足够大，则分裂形成的 g 个单簇的单簇密度 $\{den(s_1),\cdots,den(s_g)\}$ 的均值收敛于单簇 s 的单簇密度。

证明：若单簇 s 稳定，则其相应的高斯分量 $p(x)$ 可以很好地表达该单簇的样本分布。

假设分裂形成 g 个单簇 $\{s_1,\cdots,s_g\}$ 且 g 足够大，则每个单簇所在的分裂区域 R_i 足够小，使得密度函数 $p(x)$ 在区域 R_i 中几乎没有变化，则一个样本 x 落入区域 R_i 中的概率可以近似为：

$$P=\int_R p(x)\mathrm{d}x=p(x)V$$

式中，V 表示区域 R_i 的体积。假设单簇 s 的 n 个样本都是根据密度函数 $p(x)$ 独立同分布地从中抽取得到的，则其中 r 个样本落入区域 R_i 的概率为：$P_r=C_n^r P^r(1-P)^{n-r}$，根据二项分布的性质有，r 的期望 $E[r]=nP$。当样本数目 n 较大时可以用比值 r/n 作为概率 P 的估计。综上可得密度 $p(x)$ 的估计：$\hat{p}(x)=P/V=r/(nV)$。如果将单簇所在区域 R_i 定义为一个边长为 σ_i 的 d 维的超立方体，则 $\hat{p}(x)=r/(n\sigma_i^d)$，从而 g 个单簇 $\{s_1,\cdots,s_g\}$ 的概率密度可以近似为：$\hat{p}_i=\|s_i\|/(n\sigma_i^d)$，$i=1\cdots g$。由于 g 足够大，则不妨假设 $\sigma_1^d=\cdots=\sigma_g^d=\sigma^d/g$，则单簇 s 的平均概率密度也可以近似为：

$$\hat{p}=\frac{1}{g}\sum_{i=1}^{g}\hat{p}_i=\frac{1}{g}\sum_{i=1}^{g}\|s_i\|/(n\sigma_i^d)=\|s\|/(n\sigma^d)=den(s)/n$$

$$\Rightarrow\lim_{g\to\infty}\frac{1}{g}\sum_{i=1}^{g}\|s_i\|/\sigma_i^d=\lim_{g\to\infty}\frac{1}{g}\sum_{i=1}^{g}den(s_i)=den(s)$$

根据这个结论，可以对单簇的簇内稳定性进行评估。对于稳定单簇，其相应的高斯分量 $p(x)$ 可以很好地表达该单簇的样本分布，而分裂形成的单簇密度符合上述的分布特征。对于不稳定单簇，分析其形成原因：单簇中存在局部的样本聚集、存在异常点，这两种原因造成成样本点在单簇中的某些局部比较密集，而某些局部比较稀疏，即单簇的局部密度非常不均衡。根据是否存在异常点和局部样本聚集，可以将单簇分为 5 种类型，如表 1 所列。其中，无异常点/单一类簇的情况对应于稳定单簇，其余 3 类单簇均是不稳定单簇。此外，还存在一种情况：单簇中只有异常点，称这种单簇为异常单簇，其特征表现为样本点数非常少，单簇密度异常小。

表 1　4 种类型的单簇及其特性

单簇 s 的类型		密度分布	分裂形成多个单簇密度与 s 的单簇密度 den(s) 比较
存在异常点	形成的类簇		
无	单一类簇	无偏斜	均值与原单簇密度 den(s) 相近，如图 2 所示。
无	多个类簇	有偏斜	单簇密度普遍大于 den(s)，如图 3(a)所示。
有	单一类簇	有偏斜	少量单簇密度异常小，其余的均值接近 den(s)，如图 3(b)所示。
有	多个类簇	有偏斜	多数大于 den(s)，少量单簇密度异常小，如图 3(c)所示。

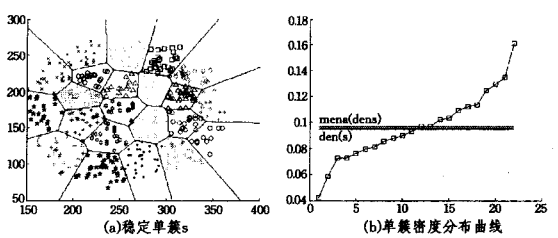

图 2　稳定单簇的密度分布曲线

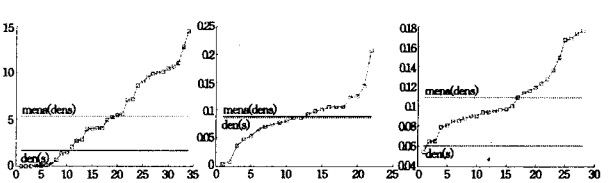

图 3　3 种类型单簇的密度分布曲线

于是，簇内稳定性评估问题转换为对非稳定单簇的检测。其中，非稳定单簇包括 3 类情况：无异常/多类簇聚集、有异常/单类簇聚集、有异常/多类簇聚集。因此，对单簇 s 的簇内稳定性，具体的检测过程包括 3 个步骤：

(1)对该单簇 s 进行聚类分裂形成 g 个单簇$\{s_1,\cdots,s_g\}$，其单簇密度为$\{den(s_1),\cdots,den(s_g)\}$；

(2)检测 g 个单簇中密度异常小的异常单簇，并将其排除在算法迭代之外，而在后处理环节对其做特殊处理。从图 3(b)和图 3(c)可以直观地看到，从异常单簇的密度到其余单簇的密度之间在曲线模式上存在剧烈的跃迁，通过检测该跃迁可以确定一个密度阈值 $denh$，则单簇密度小于 $denh$ 的单簇即为异常单簇。

(3)对除异常单簇外的其它单簇，检测其单簇密度是否普遍大于 $den(s)$，如图 4(a)和图 4(c)所示，若是则说明存在类簇聚集，因此需要分裂。

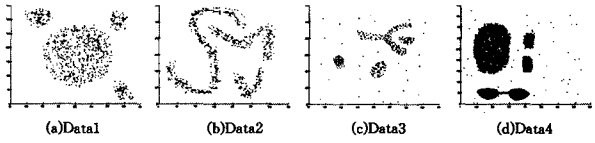

图 4　常用的评估聚类算法的模拟数据

簇间稳定性评估是检测两个单簇之间是否具有很强的归并于同一个类别的倾向，其评估的指标称为簇间融合性。

定义 2(簇间的融合性)　假设两个单簇 s_1,s_2，其对应的高斯分量为 g_1,g_2，密度参数为(μ_i,σ_i)，$i=1,2$，则二单簇之间的融合性定义为：

$$M(s_1,s_2)=\frac{\|\{x\mid d(x,\mu_1)<2\sigma_1 \wedge d(x,\mu_2)<2\sigma_2\}\|}{\|\{x\mid d(x,\mu_1)<2\sigma_1 \vee d(x,\mu_2)<2\sigma_2\}\|} \quad (4)$$

式中，$d(x,\mu_i)$表示样本点 x 到 g_i 的均值中心的距离，$\|*\|$ 表示集合 $*$ 的基数。根据球形高斯分布的性质，随机变量的取值 95%都落在以均值为中心半径为 2σ 的球形区间内，这里两个单簇之间的融合性是指同时处于两个高斯分量的 $2\sigma_i$ 球形区间内的样本点比例。自然地，该比例越大则说明合并倾向越大。

定义 3(类簇与聚类结构)　类簇 C 是由一组非异常的单簇组成的集合，对于该集合中任意两个单簇 s_1 和 s_2，或者 $M(s_1,s_2)$大于给定阈值，或者存在一组单簇 $s_{i1}=s_1,s_{i2},\cdots,s_{ik}=s_2$，使如下取值均大于给定阈值。

$\{M(s_{ij},s_{i(j+1)})\mid j=1,\cdots,k-1\}$

而聚类结构 R 是由一组类簇组成的集合。

在分裂与合并过程所产生的聚类结构的基础上，经过后处理环节，即对异常单簇中的样本进行甄别判断，生成最终的聚类结果。

3　SMClus 算法描述

SMClus 算法是一种迭代增强聚类结构稳定性的聚类过程，具体包括 3 个阶段：

(S1.) IntraStabEval-Split：簇内稳定性评估与分裂阶段，生成一系列稳定单簇；

(S2.) InterStabEval-Merge：簇间融合性评估与合并阶段，生成一系列类簇，组成聚类结构；

(S3.) PostProcessing：后处理阶段，产生最终聚类结果。

算法初始假设只有一个单簇 s，且包含所有样本点：

算法 S1　IntraStabEval-Split：簇内稳定性评估与分裂

```
输入：包含所有样本点的单簇 s;
输出：一组稳定单簇 SetOfStableSngCluster 和异常单簇 SetOfOutlierSngCluster;
Begin
  //待判定的单簇集合，初始时只有单簇 s 一个元素
  Var SetOfSngCluster = {s};
  //稳定单簇集合，初始为空
  Var SetOfStableSngCluster = {};
  //异常单簇集合，初始为空
  Var SetOfOutlierSngCluster = {};
  Do {
    SngCluster = SetOfSngCluster.pop();//得到单簇
    /* 判断单簇 SngCluster 是否稳定 */
    SplittedSngClusters = split(SngCluster);
    dens = calcDensity(SplittedSngClusters);
    //剔除异常单簇
    OutlierSngCluster = deOutlier(SplittedSngClusters);
    SetOfOutlierSngCluster.pushback(OutlierSngCluster);
    bStable = isStable(dens);
    //若单簇 SngCluster 稳定，则将其加入稳定单簇集合，否则将其二分后加入待判定单簇集合
    If (bStable) {
      SetOfStableSngCluster.pushback(SngCluster);
    } Else {
      TwoSgnClusters = biSplit(SngCluster);
      SetOfSngCluster.pushback(TwoSgnClusters);
    }
  } While(! SetOfSngCluster.isEmpty());
End
```

在算法 IntraStabEval-Split 中，首先定义了 3 个集合：有待判定稳定性的单簇集合 SetOfSngCluster(初始只包含单簇 s)、判定为稳定的单簇集合 SetOfStableSngCluster(初始为空)、判定为异常的单簇集合 SetOfOutlierSngCluster(初始为空)。然后，针对待判定单簇集合中的每个单簇，判定其稳定性，若稳定则将其加入稳定单簇集合，否则将其二分(biSplit)并加入待判定的单簇集合。对单簇 SngCluster，判断其稳定性的具体流程包括：首先通过 *split* 函数对单簇 SngCluster 实施分裂，得到一系列单簇 SplittedSngClusters，然后计算 SplittedSngClusters 中各单簇的密度，根据密度序列剔除异常单簇并将异常单簇加入集合 SetOfOutlierSngCluster，最后根据剩余非异常单簇的密度序列进行稳定性评估(isStable)。

簇内稳定性评估与分裂阶段完成后，会生成一系列稳定单簇。接下来要对这组稳定单簇评估其簇间融合性，并进行合并处理。

算法 S2　InterStabEval-Merge：簇间融合性评估与合并

```
输入：稳定单簇集合 SetOfStableSngCluster;
输出：生成聚类结构 R;
Begin
  //根据式(4)计算两两稳定单簇之间的融合性
  mergs = calcMergs(SetOfStableSngCluster);
  //融合阈值确定，通常推荐 mergh = 0
  mergh = thresholding(mergs);
  //根据融合阈值，合并稳定单簇
  R=merge(SetOfStableSngCluster,mergh);
End
```

后处理过程就是根据稳定的聚类结构将异常类簇中的异

常样本识别出来，并将正常样本赋予某个类簇的标记，最终得到聚类结果。基本思路是给定稳定聚类结构 R 和异常单簇 s，对任意样本点 $x \in s$，设 R 中到 x 距离最小的类簇为 C，则称 x 为异常点，如果有：

$Dist(x,C) > 2.698 * sigma(s_i)$，$s_i \in C$ 且满足 $Dist(x, mean(s_i)) = Dist(x,C)$

式中，$sigma(s_i)$ 表示该单簇的球形高斯分量的标准差。此外，由于每个单簇均由一个高斯分量进行描述，其理论的上截断点是 2.698 倍的标准差。类簇 C 的代表结构定义为每个单簇的中心点组成的点集，而样本点 x 到类簇 C 的距离定义为 x 到 C 的代表点集的最小距离。

算法 S3 后处理算法：PostProcessing

```
输入：聚类结构 R，异常单簇集合 OulierSngClusters；
输出：最终的聚类结果；
Begin
  For Each x In OulierSngClusters Do
    //计算到 x 最近的单簇 C 及其所属类簇 si，
    // 以及最近距离 dist
    [C,si,dist] = CalcMinDist (R,x)；
    If dist > 2.698 * sigma(si) Then x 为异常点；
    Else 将 x 赋予类簇 C；
  End Do
End
```

4 仿真实验

本节通过仿真实验对 SMClus 算法的聚类结果的有效性进行分析。实验采用的数据包括人工构造的二维模拟数据（如图 4 所示）和多光谱遥感图像数据。对于二维模拟数据，直接通过视觉判断和分类精度判断结果的有效性；对于真实的多光谱遥感图像数据，通过已有的像素类别标签计算分类精度来判断聚类结果的有效性。实验的硬件环境为 P4 1.86 GHz 的 CPU 和 512MB 的内存，软件环境为 Microsoft Windows XP professional 操作系统，所有代码均在 Matlab7.04 下实现。

图 5 和图 6 分别给出了上述 Data1 和 Data2 两组数据经过分裂、合并后的结果，以及相应的 DBSCAN 聚类结果。实验中，单簇数目下界 $lbound\_nclus = 30$，样本点数目下界 $lbound\_nsamp = 5$，合并阈值 $mergh = 0$。在每个图中，图(a)是经过簇内稳定性评估与分裂阶段后的结果，图(b)是经过簇间融合性评估与合并阶段后的结果，图(c)是 DBSCAN 算法的聚类结果。

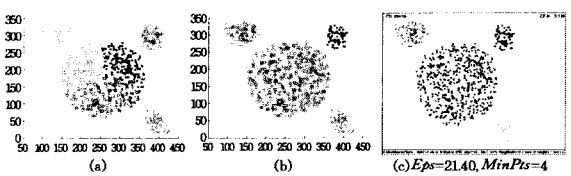

图 5 模拟数据 Data1 的聚类处理结果

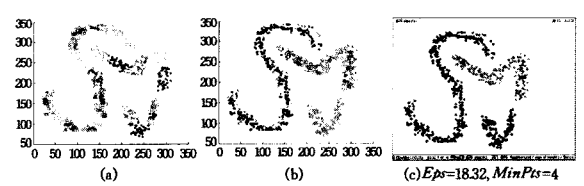

图 6 模拟数据 Data2 的聚类处理结果

可以看出，SMClus 算法对上述 4 组数据无须后处理步骤即可得到较高质量的聚类结果。DBSCAN 算法亦可取得不错的结果，但其密度参数设置需要多次调试，若用户对数据没有先验知识，则难以应用。

真实数据采用从 http://archive.ics.uci.edu/ml/datasets.html 站点下载来的 Statlog（Landsat Satellite）数据集[10]，是从多光谱 Landsat 卫星遥感图像数据中采样生成的，包含 6435 个样本点，每个样本包含 37 个属性，其中前 36 个属性是一个 3×3 邻域中 9 个像素的光谱值，最后一个属性是中心像素的类别标记，表示中心像素所处位置的土壤覆盖类型，共 6 类（red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, very damp grey soil）。具体实验采用中心像素的 4 个光谱值，聚类结果通过 F 度量来评价，其度量公式如下：

$$F(i,j) = \frac{2 \times recall(i,j) \times precision(i,j)}{recall(i,j) + precision(i,j)}$$

式中，$recall(i,j)$ 为召回率，$precision(i,j)$ 为准确率，分别定义为：

$$precision(i,j) = n_{ij}/n_j,\ recall(i,j) = n_{ij}/n_i$$

式中，n_i 和 n_j 分别代表第 i 类和第 j 类中样本的数目，而 n_{ij} 则是被分到类 j 中的类 i 对象的数目。具体的实验结果如表 2 所列。

表 2 Reuters-21578 数据集上的实验结果（h=10 * k）

数据集	文档数/类别数 k	DBSCAN F 值	SMClus F 值
Statlog_1	1000/3	66.3%	74.5%
Statlog_2	2000/4	69.8%	77.8%
Statlog_3	4000/5	68.4%	75.7%
Statlog_4	6435/6	61.5%	82.0%

从表 2 可以看出，对真实数据而言，SMClus 算法聚类效果较 DBSCAN 算法有大幅度的提升，且无需过多的人工干预。当然，SMClus 算法的时间开销则增加了很多。

结束语 本文提出一种面向结构稳定性的迭代聚类方法框架，并在此基础上设计并实现了 SMClus 聚类算法，综合考虑了假设结构的有效性、算法有效性，同时在很大程度上消除了聚类算法对先验知识有效性的依赖。从实验结果可以看出，SMClus 算法在聚类结果的有效性上有了较大提升。然而，SMClus 算法会消耗过多的计算资源，这是下一步改进的重点。此外，SMClus 中采用的密度估计方法限制了算法在较高维度的数据集上的适用性，这也是有待考虑的问题。

参 考 文 献

[1] Han J W, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco, CA: Morgan Kaufmann Publishers, 2001: 223-250

[2] David M. Information Theory, Inference and Learning Algorithms[M]. Cambridge University Press, 2003: 284-292

[3] Zhang T, Ramakrishnan R, Linvy M. BIRCH: An efficient data clustering method for very large databases[C]// Jagadish HV, Mumick IS, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Montreal: ACM Press, 1996: 103-114

[4] Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases[C]// Haas LM, Tiwary A, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1998: 73-84

[5] Karypis G, Han EH, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling[J]. Computer, 1999, 32(8): 68-75

[6] Ester M, Kriegel HP, Sander J, et al. A density-based algorithm for discovering clusters in large spatial database with noise[C]// Simoudis E, Han J, Fayyad UM, eds. Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996: 226-231

[7] Ankerst M, Breuning M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure[C]// Delis A, Faloutsos C, Ghandeharizadeh S, eds. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Philadelphia: ACM Press, 1999: 49-60

[8] Polikar R. Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features[J]. IEEE Signal Processing Magazine, 2007, 24(4): 59-72

[9] Kuncheva L I. Combining Pattern Classifiers, Methods and Algorithms[M]. New York, NY: Wiley Interscience, 2005

[10] Asuncion A, Newman D J. UCI Machine Learning Repository [OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, School of Information and Computer Science, 2007

[11] 雷小锋，谢昆青，林帆，等. 一种基于 KMeans 局部最优性的高效聚类算法[J]. 软件学报，2008，19(7)：1683-1692

[12] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61

(上接第 213 页)

图 2 为 0dB 的带噪语音经不同方法增强后的时域波形图。从图中不难发现，对含噪语音在使用本文方法增强后，比采用基本谱减法、文献[2]中提出的方法增强后都有明显的改善。同时，本文对不同信噪比条件下增强前后的语音进行主观试听实验。试验结果表明，采用基本谱减法对带噪语音进行增强后虽然噪声已明显减少，但又产生了有节奏的音乐噪声；采用文献[2]中提出的方法对带噪语音进行增强后，不仅噪声明显减少，而且音乐噪声也得到了一定的抑制；采用本文方法对带噪语音进行增强后，效果较基本谱减法有了很大改善，较文献[2]中提出的方法也有一定程度的改善，尤其在低信噪比时较为明显，所得结果更加容易让人接受。

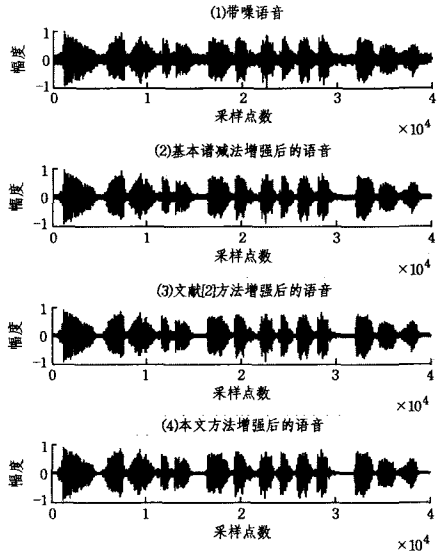

图 2 带噪语音和 3 种不同方法增强后的语音

客观评价选择信噪比(SNR)作为衡量指标，信噪比的定义为

$$SNR=10\lg\frac{\sum_{i=1}^{n}[s(i)]^2}{\sum_{i=1}^{n}[y(i)-s(i)]^2} \tag{14}$$

式中，n 为采样点数，$s(n)$为增强后的语音，$y(n)$为带噪语音。对输入信噪比为−5dB，0dB，5dB 的带噪语音分别采用基本谱减法、文献[2]中提出的方法以及本文方法进行语音增强。各增强方法的输出信噪比如表 1 所列。

表 1 3 种增强方法输出信噪比对比表 (单位：dB)

输入信噪比	基本谱减法输出信噪比	文献[2]方法输出信噪比	本文方法输出信噪比
−5	2.85	3.94	8.12
0	5.76	7.37	10.95
5	8.92	11.06	13.26

从表 1 中可以看出，本文方法的增强效果优于基本谱减法和文献[2]中提出的方法，噪声被更好地消除，信噪比得到了进一步提高。尤其是输入为低信噪比语音时，本文方法对信噪比的改善效果更加明显。

结束语 语音增强是语音信号处理的前沿领域，也是语音识别和语音合成等方向的基础。目前已存在众多的针对平稳噪声环境下的语音增强技术。然而，许多环境下的噪声都是非平稳的，本文提出了基于实时噪声估计的改进谱减法。通过与其他方法比较可知，本文提出的方法具有良好的降噪性能，较大程度地抑制了音乐噪声，减少了语音失真，提高了语音质量，特别是对于信噪比较低的情况，笔者方法优势明显。

参 考 文 献

[1] Yamauchi J, Shimamura T. Nonstationary Noise Estimation Using High Frequency Regions for Spectral Subtraction[J]. IEICE Transactions on Communications, 1998, 70(3): 335-349

[2] Yamashita K, Shimamura T. Nonstationary Noise Estimation Using Low-frequency Regions for Spectral Subtraction[J]. IEEE Signal Processing Letters, 2005, 12(6): 465-468

[3] Virag N. Single Channel Speech Enhancement Based on Masking Properties of Human Auditory System[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(2): 126-137

[4] Cohen I, Loizou P. Speech enhancement based on wavelet th - resholding and multitaper spectrum[J]. IEEE Signal Processing Letters, 2002, 9(1): 12-15

[5] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Transactions on Speech and Audio Processing, 2001, 9(5): 504-512

[6] 陈国明，赵力，邹采荣. 一种基于短时谱估计和人耳掩蔽效应的语音增强算法[J]. 电子与信息学报，2007，29(4)：863-866

[7] 武文娟，顾宏斌，潘秀林. 基于临界带特征矢量距离的端点检测算法[J]. 计算机科学，2009，36(2)：220-221