聚类方法综述

金建国
(浙江工业大学理学院应用数学系 杭州 310032)

摘 要 文中对聚类方法作了综述。系统地讨论了聚类研究中的4个关键内容:数据点之间"距离"函数的定义方法、类数目的确定、高效优良的聚类算法和聚类算法好坏的评估。分析了各类聚类算法的优劣性,指出了聚类分析研究的发展趋势。

关键词 聚类,距离,类数目,算法评估
中图法分类号 TP18 **文献标识码** J

Review of Clustering Method

JIN Jian-guo
(Applied Mathematics Department of Science College, Zhejiang University of Technology, Hangzhou 310032, China)

Abstract The paper reviewed some clustering methods and results. Four key problems were discussed: distance and similarity measures, cluster number, clustering algorithms and the valid methods. The advantages and disadvantages of clustering algorithms were analyzed. The developing trend of clustering analysis techniques was pointed out.

Keywords Clustering, Distance, Cluster number, Algorithm valid

1 引言

由聚类所生成的簇是一组数据对象的集合,这些在同一个簇中的对象彼此相似,而与其他簇中的对象相异。聚类分析最大程度地实现类中对象相似度最大、类间对象相似度最小。

聚类算法大体可以划分为以下几类[21,40,41]:(1)基于划分的方法(partitioning method)[2,3,7,9,13,14,20,39];(2)基于层次的方法(hierarchical method)[11,12,15];(3)基于密度的方法(density-based method)[5,6,17,18];(4)基于网格的方法(grid-based method)[19];(5)基于模型的方法(model-based method)[10,28,29];(6)模糊聚类方法(fuzzy method)[26];(7)基于图论的方法[16];(8)基于分形的方法[40];(9)复杂网络聚类方法[41];(10)仿生法[38];(11)核聚类方法[42]。不少聚类算法是这些方法的综合。

聚类算法在工程(如机器学习、人工智能、机械工程、电子工程)、计算机科学、生医学、地球科学(如遥感)、社会科学(如心理学)、经济学等领域都有广泛的应用。聚类算法和模式识别、模式分类密切相关,是模式识别和模式分类的基础。聚类算法可直接应用于二维多边形剖分[1,36]、三维网格有意义剖分[9]和图像分割[37]中,为二维多边形物体、三维网格物体和图像的识别打下了基础。

直接导致数据集聚类结果好坏的因素有3个[22-27,30]:类数目的正确获得、决定数据点之间亲密度的距离函数和高效优良的聚类算法。因此评价聚类算法的优劣也成为目前聚类研究的热点之一。关于聚类方面的研究,主要集中在类数目的确定、数据点间亲密度(距离函数)的定义、开发优良的聚类算法和对聚类算法及聚类结果好坏的评估这4点上。本文结合我们在多边形和网格有意义剖分的工作[1],对聚类算法、距离函数、类数目的确定和算法评估这4个内容进行了系统的论述。

2 聚类算法概述

2.1 聚类算法

2.1.1 基于划分的方法

主要有K-Means[2-4]、K-modes[14]、PAM(Partitioning Around Medoids)[20]、CLARA(Clustering Large Applications)[13]、AP(Affinity Propagation Clustering)[7]、SPEA(Spectral Analysis)[9,39]等聚类算法。

最经典的聚类算法是K-Means,它由MacQueen于1967年提出。其主要思想是找出数据集的k个类中心(质心),把数据集划分为k个类,使得数据集中的数据点与所属类的类中心的距离平方和最小。该算法对初值敏感,需人工指定类数k,优点是算法简单易于实现。K-modes[14]是K-Means算法的一个延伸,主要是可处理分类属性数据(categorical data),而不像K-Means那样只能处理数值属性的数据(numerical data)。K-Means和K-modes都不能处理孤立点(outliers)情形。PAM(Partitioning Around Medoids)[20]以数据集中的实际数据点为Medoids进行聚类,而不像k-means中以质心(未必恰好是数据集中的数据点)为类中心。PAM可处理孤立点等奇异情形,但该算法计算量非常大。

CLARA[13]和PAM方法相似,主要是为了减少PAM中

本文受浙江省自然科学基金(Y1100837),浙江省151人才培养计划资助。
金建国(1970-),男,博士,副教授,主要研究方向为计算机图形学、模式识别与模式分类,E-mail:npy20022003@zjut.edu.cn。

的计算量。该算法先从数据集中提取一部分数据点作为样本，然后对样本采用 PAM 算法聚类，利用这样的一个思想可大大减少计算量。

亲密度传播聚类(Affinity Propagation clustering，AP)[7]是 Frey 等人 2007 年提出的一种聚类算法，该算法快速、有效。AP 算法经过人脸图像聚类、文本中关键句子选择、基因片段聚类以及航空路线规划等试验中的测试，证明其不仅聚类结果更好，而且对于大规模数据集而言，花费时间只有其它聚类方法的百分之一[7]。对于小规模的数据集，该算法聚类结果的正确性与效率与其它方法相当或略优，偶尔甚至不及[8]。AP 算法初始时将所有的数据点都视为潜在的聚类中心，将两个点之间的欧氏距离的负值设想为吸引度或归属度，则点 k 对较近的点 i 的吸引力比较大，同样点 i 认同点 k 为其聚类中心的归属感也较强。这样，数据点 k 对其他数据点的吸引力之和越大，成为聚类中心的可能性也越大，反之可能性就越小。以此原理出发，AP 算法为选出合适的类代表(类中心)而不断从数据点集中搜集和传递有关的消息(Message)：为候选的类中心点 k 从数据集中每个数据点 i 搜集消息 $R(i,k)$(称为点 k 对点 i 的 responsibility 或吸引度)来描述数据点 k 作为数据点 i 的类中心的适合程度；同时收集消息 $A(i,k)$(称为点 i 对点 k 的 availability 或归属度)来描述数据点 i 选择数据点 k 作为其类中心的适合程度。$R(i,k)$ 与 $A(i,k)$ 越大，点 k 作为最终聚类中心的可能性就越大。AP 算法通过迭代，循环不断地进行消息的搜集和传递，以产生 m 个高质量的类中心和对应的聚类，同时聚类的目标函数也得到了最优化，数据集中各个数据点也最终归队于各个以类中心为代表的所属的类。AP 算法并不明确要求指定类数目。但我们在研究该算法并实际应用时发现，有时对亲密度矩阵的一个微小扰动就会影响其聚类结果，并且有时会出现两个数据点互为类中心的情况。

AP 算法与 K-means 算法等同属于 K 中心聚类方法。经典的 K-means 算法的优点是简单、快速而且能有效处理大规模数据集，然而算法对初始聚类中心的选择敏感且容易陷入局部极值，因此需要在不同初始化下运行很多次，以寻找一个相对较好的聚类结果。但这种策略也只有在非海量数据和较小的类数及某次初始化靠近好的结果时才有效。另外，它要求用户必须事先给出聚类个数 k。AP 算法部分地克服了这些缺点，其迭代过程不断搜索合适的聚类中心，同时使得聚类的目标函数 $E(C)$ 最优化。若各个类的结构比较紧密，算法则容易保证各个类的亲密度和均比较大，从而能给出比较正确的聚类结果；但对于比较松散的聚类结构，算法倾向于产生较多的类来实现 $E(C)$ 最大化，这使得算法产生的聚类类数过多，而不能给出准确的聚类结果。这种不足在很大程度上会限制其应用范围。

谱分析方法(Spectral Analysis)[9,39]利用特征值和特征向量的方法对数据集中的数据点进行聚类，取得了较好的结果，在聚类和模式分类、模式识别、网格剖分(Mesh Segmentation)中得到了广泛的应用。和 AP 方法一样，它同样是对数据点集的亲密度矩阵 $\boldsymbol{S}$ 进行分析，但聚类时不是直接采用 $\boldsymbol{S}$ 矩阵，而是先计算它的 k 个最大的特征值和特征向量，并利用它们构建一个对称矩阵 $\boldsymbol{Q}$，进而对 $\boldsymbol{Q}$ 进行亲密度分析，得到数据集的最终聚类。谱分析方法并不是直接对亲密度矩阵进行分析而得出分类结果，由于它要先计算特征向量，因此计算量较大，并且也同样需要由用户事先给出聚类个数 k。谱分析方法得到的包含数据点亲密度信息的矩阵 $\boldsymbol{Q}$ 与原始亲密度矩阵 $\boldsymbol{S}$ 相比，亲密度信息 $\boldsymbol{Q}$ 只是 $\boldsymbol{S}$ 的一个近似，近似的程度和值 k 相关，k 取得越大，$\boldsymbol{Q}$ 和 $\boldsymbol{S}$ 的误差越小。k 是用户指定的数据点集最终分类结果的类数，是算法的一个参数。$\boldsymbol{Q}$ 和 $\boldsymbol{S}$ 相比的优势在于，对 $\boldsymbol{Q}$ 进行亲密度分析实现分类要比对 $\boldsymbol{S}$ 直接进行分析实现分类要容易得多，这在极化理论中得到了证明。我们利用谱分析方法研究了多边形与网格有意义的剖分，并在此基础上开发了一个原型系统[1]，取得了良好的剖分结果。

2.1.2 基于模型的方法

基于模型的方法[10,28,29]主要可以分为两类：一类是利用混合概率密度分布模型(Mixture Models)来聚类[28,29]，另一类是利用统计物理学中的非均匀铁磁模型(inhomogeneous ferromagnetic model)的顺磁阶段来聚类[10]。

混合概率密度分布模型聚类法用混合概率密度函数来拟合数据集，令混合概率密度函数：

$$f_{mix}=\sum_{k=1}^{G}\tau_k f_k(y|\theta_k)$$

其中，$\tau_k\geqslant 0,\sum_{k=1}^{G}\tau_k=1$，$f_k$ 和 θ_k 分别是混合模型中第 k 个组成部分的密度函数和参数，τ_k 是数据集中数据点 $\boldsymbol{y}$ 属于第 k 个组成部分(第 k 个 cluster)的概率。设 $\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$ 是需分类的数据集，令 $\boldsymbol{x}_i=(\boldsymbol{y}_i,\boldsymbol{z}_i)$，$i=\overline{1,n}$，称之为完全数据(complete data)，其中 $\boldsymbol{z}_i=(z_{i1},z_{i2},\cdots,z_{iG})$ 是未知部分，$z_{ik}=\begin{cases}1, & \text{若 } \boldsymbol{x}_i \text{ 属于类 } k\\ 0, & \text{否则}\end{cases}$，$k=1,2,\cdots,G$。则对应于“完全数据”$\boldsymbol{x}_i$ 的对数似然函数，利用著名的 EM 算法[28]可迭代计算出 z_{ik}，τ_k 及 θ_k 等未知参数的近似值，从而由 z_{ik} 的值获得数据集 $\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n$ 的一个分类。混合模型的组成部分密度函数可根据需要选择，较为常用的有 Poisson 分布、高斯分布、t 分布等。成份数 G(即数据集的类数)通常可用模型选择的方法来进行选择，通过计算不同模型所得到的 AIC 值(Akaike Information Criteria)和 BIC 值(Bayesian Information Criterion)[29]加以筛选。若组成成份采用高斯分布，则其参数中的均值 μ_k(为第 k 类的类中心)、协方差矩阵 $\boldsymbol{\sigma}_k$(揭示了第 k 类的形状和方位)均可由 EM 算法计算。该算法由于采用了 EM 算法，因此聚类结果对初值敏感。

顺磁聚类法[10]利用统计物理学方法对聚类作了研究，主要利用了非均匀铁磁模型(inhomogeneous ferromagnetic model)的物理性质。对每一个数据点赋以一个自旋方向，规定只有最相邻的数据点之间才会对彼此的自旋方向相互影响。最近邻间的相互影响用自旋关联(spin-spin correlations)函数值来表示，这个模型可用著名的伊辛模型(Ising model)或 Potts 模型描述。自旋关联值通常用蒙特卡洛方法模拟获得。当铁磁温度由绝对零度逐步升高到达居里温度时，数据集(铁磁模型)由铁磁性阶段(ferromagnetic phase)过渡到顺磁性阶段，这时铁原子(数据点)的自旋方向由完全有序一致变成不同区域局部一致状态，数据集的 clusters 就显现。算法利用各个数据点处的自旋关联值，在顺磁性阶段(温度 T 在某个范围时)确定所有的 cluster。算法的关键点是确定顺磁状态临界温度以及用蒙特卡洛方法模拟计算自旋关联值。

算法的特色是把数据点间的距离函数转化为自旋关联值，从而可处理在密度低的区域两个相邻点虽然距离很小但属于不同类而在高密区域同样距离却属于同类的情形。另外算法稳定，对初值不敏感。

2.1.3 基于分形的方法

基于分形的方法如 FC(Fractal Clustering)[40,43] 等主要从分形维数着手来完成聚类。FC 算法的基本思想是认为在同一个 cluster 内部的任何一个数据点的改变都不太可能引起该 cluster 原有分形维数的本质变化。FC 首先采用网格的聚类算法对数据集的一个样本集进行初始聚类，初始得到的每一个 cluster 要保证有足够多的数据点，以能够计算该 cluster 的分形维数。然后对数据集中未归类的每一个数据点 p，计算其插入每一 cluster 后引起的该 cluster 分形维数的变化值(绝对值)，若所有 cluster 变化值的最小值大于某一设定的阈值，则认为此数据点为噪声，去除。否则把 p 点归类于引起 cluster 分形维数变化最小的那个类。该算法能处理噪声，可处理任意形状的 cluster，能处理高维数据集。

2.1.4 模糊聚类法

在实际应用中，数据集中的数据点有时并不仅仅只是属于某一类的，而是同时属于多个类。为处理这个问题，模糊聚类法[26,44,45] 如 FCM 算法、FBSA 算法、Gustafson-Kessel 算法、Gath-Geva 算法等就应运而生。其中最有代表性的当属 FCM(Fuzzy C-means)算法[26]。FCM 的基本思想是使得目标函数 $J_m(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}^{m}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_A^2$ 取得最小值，其中 $\boldsymbol{U}=(u_{ki})_{n\times c}$ 为模糊剖分矩阵，u_{ki} 表示数据点 $\boldsymbol{x}_k$ 属于类 i 的程度，满足 $\sum_{i=1}^{c}u_{ki}=1$ 且 $u_{ki}\geqslant 0$。m 是参数，称之为模糊指数(fuzzifier)。用迭代法计算模糊剖分矩阵 $\boldsymbol{U}$ 和类中心 $\boldsymbol{V}=\{\boldsymbol{v}_i\mid i=\overline{1,c}\}$，并由此获得类中心和分类结果。K-means 算法、Fuzzy C-means(FCM)算法、Gustafson-Kessel 算法和 Gath-Geva 算法最优化的目标函数中只是距离 $\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_A^2$ 的定义不同，却导致了它们处理不同 cluster 形状的能力有强弱。Gath-Geva 算法和 Gustafson Kessel 算法比 Fuzzy C-means 算法、k-means 算法能处理的 cluster 形状更加丰富。

2.1.5 其它聚类算法

比较有代表性的主要有：基于层次的 CURE[11]、ROCK[12] 和 BIRCH[15]；基于密度的 FDBA[5]、bDBSCAN[6]、DBSCAN[17] 和 ST-DBSCAN[18]；基于网格的 STING[19]；基于图论的 CLICK[16]；复杂网络聚类法[41]；仿生法[38] 以及核聚类方法[42]。以下对每一类型选其一代表性算法作简要阐述。

CURE(Clustering Using Representatives)是个层次聚类算法。其主要思想是在层次聚类的两个类合并过程中，在合并的 cluster 里适当选取分布较散的一些样本点(selecting well scattered points)，然后样本点按照设定的收缩率 $\alpha\in[0,1]$ 向该 cluster 中心收缩后获得该 cluster 的代表点集(representatives)，算法在下一层次时考虑某两个 cluster 是否合并就以它们两者的这些代表点集是否距离最近作为依据。该算法能很好地处理孤立点问题，并且能处理各种形状的 cluster，克服了一些聚类算法只能处理圆形或球形的 cluster 这个问题。基于密度的 DBSCAN(Density Based Clustering Algorithm)算法的主要思想是一个 cluster 中的每一个点在给定半径的邻域内必须至少含有某个给定数目的点，因此它能处理孤立点，并且只需一个参数值。该算法理论上能处理除两个 cluster 之间有致密点集相连(哑铃状)的任意形状的 cluster。基于网格的 STING(Statistical Information Grid)用层次结构的方式把数据空间划分为很多个矩形单元(或网格)，然后计算网格里数据点的统计值(包括均值、标准差、最大小值、分布类型等)，并利用这些信息进行聚类。该方法利用层次结构的矩形单元存储法(下层的单元是上层单元的子单元)，能较快地进行信息查询从而减少算法的计算量。基于图论的 CLICK(Cluster Identification via Connectivity Kernels)递归地对图进行最小权重分割(minimum weight cut)来产生 clusters，并假设 cluster 内部和 clusters 之间的相似度服从不同均值和方差的高斯分布，均值与方差这些参数值通过最大似然估计方法或 EM 算法计算获得。算法主要过程如下：对一个图，首先用递归最小权重分割法得到各个 cluster 的内核(kernel)，内核中的每一个数据点均属于该 cluster。在递归分割过程中，未进入某个 kernel 的数据点则进入单点集(the singleton set)R，然后通过 singtons 向 kernels 的归队及 kernels 的合并处理等循环过程来最终获得数据集的聚类。CLICK 算法的优点是速度快，聚类正确度较高。

复杂网络聚类法[41] 用于发现网络簇结构，在社会网、生物网和万维网中有着广泛的应用。例如 Kleinberg 提出的 HITS 算法，利用 WWW 中 authority 和 hub 两种基本页面的引用关系发现由 authority-hub 构成的网络簇结构，算法被广泛应用于多个搜索引擎中[41]。网络簇结构是复杂网络最普遍和最重要的拓扑结构属性之一，具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点。复杂网络聚类的具体算法已有很多，如基于优化的网络聚类算法 Kernighan-Lin 算法、快速 Newman 算法和 Guimera-Amaral 算法以及基于启发式的网络聚类算法如 MFC 算法和 HITS 算法。对于已知簇结构的随机网络模型，基于优化的聚类算法比启发式算法有更好的聚类精度。

仿生法[38] 如人工鱼群聚类算法、蚁群聚类方法等是一种基于动物或生物行为的群体智能优化聚类算法。这些算法把人工鱼群、蚁群等经典优化算法和传统聚类算法相结合，以克服目前聚类分析算法中普遍存在的对初始参数敏感、难以找到最优聚类以及聚类有效性等问题。仿生聚类法具有良好的克服局部极值和获得全局极值的能力。

核聚类法[42] 利用 Mercer 核把输入空间的样本映射到高维特征空间后，在特征空间中进行聚类。由于经过了核函数的映射，使原来没有显现的特征突现出来，从而能够更好地聚类。核聚类方法在性能上比经典的聚类算法有较大的改进，具有更快的收敛速度以及更为准确的聚类。仿真实验的结果证实了核聚类方法的可行性和有效性[42]。该算法从某种意义上来说，和谱分析法有异曲同工之妙，目的都是放大特征的显现，使同一个 cluster 的数据点联系更加紧密，而不同 cluster 之间更分散，使之更容易聚类。要达到这个目的，关键是如何构造关于距离的映射函数。

2.2 距离的确定

数据点间亲密度或距离如何定义直接影响着聚类结果能否正确获得。对于很多数据集，用欧氏距离作为定义数据点

间亲密度的基础，即可获得较好的聚类结果。可以说欧氏距离是聚类分析中最为常见的数据点间距离定义方法（或数据点间亲密度定义的基础）。另外常见的“距离函数”定义还有以下几种（设数据点维数为 m）：皮尔森相关距离 $(1-r_{ij})/2$（其中 r_{ij} 为 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的相关系数，该距离广泛应用于基因分析）、Minkowski 距离 $(\sum_{k=1}^{m}|\boldsymbol{x}_{ik}-\boldsymbol{x}_{jk}|^{\frac{1}{n}})^n$、Mahalanobis 距离 $(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)$（其中 $\boldsymbol{M}$ 为协方差矩阵）和余弦距离 $(\cos\alpha=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)/(\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|)$，广泛应用于文本聚类）等。

距离函数的定义要具体问题具体分析，不一定要满足度量公理，如可以是广义距离，也可以是某些距离的组合。距离函数定义得是否合适，直接影响着最终的聚类结果是否正确。测地距离和角距离在聚类分析中也应用得很多，例如在 Mesh Segmentation[9] 的研究中，网格间的距离就定义为两者间测地距离与法向角距离的一个线性组合。也有学者认为用距离函数作为定义亲密度的基础是值得改进的，距离小只是表明数据点间各个分量间有较近的值，而两物体（数据点）只有展示出相关联的内在结构才能表明它们相似，数据点间的距离远并不能表示它们不相似，例如基因数据。因此提出了基于改进了亲密度定义（或距离）的 Pcluster 模型[23]。还有学者提出基于概念相似[22]（concept similarity）和 ISOMAP based metrics[30] 的亲密度定义方法。我们在多边形与网格物体有意义的剖分研究中[1]，对多边形顶点之间或 Mesh 的网格之间如何定义距离（或亲密度）作了深入的思考，主要采用了测地距离、顶点（或网格）间的可见性来解决这个问题。

当数据是高维时，为了距离函数定义的方便或分类结果的图形可视化，往往采用降维的方法[9,21,30-32]。降维法通常采用 PCA（Principal Component Analysis）方法、MDS（Multi Dimensional Scale）法、ISOMAP 法、谱分析方法、SM（Sammon Mapping）方法和投影寻踪（Projection Pursuit）法。其它还有 Wavelet transform 法[35]、Singular value decomposition 法[33] 和 *nonnegative matrix factorization* 方法[34]。用这些降维法可以找出高维数据集其真正的内在结构维数，达到降维目的。PCA 和 MDS 方法简单易于实现，但只能发现线性或拟线性子空间的真正内在结构，而 ISOMAP 却能发现 PCA 或 MDS 不能发现的数据集中的非线性结构[31]。ISOMAP 的主要思想是通过计算流形上点间的测地距离，结合 MDS 法，实现寻找非线性结构以及降维。PCA 法、ISOMAP 法和谱分析法均从特征值、特征向量着手，来完成数据集从 n 维到 q 维的嵌入。例如 PCA 的本质是把原来 n 维空间的一组坐标系换成另一组 n 维正交坐标系，使得在这组新的坐标系下，在“主要坐标轴”方向（对应于“最大”的几个特征向量方向）数据集的几何属性和结构有较强体现，并舍弃部分相对不重要的坐标轴方向（对应于特征向量“较小”的）来达到尽量不丢失原来数据几何信息而又降维的目的。Sammon Mapping 降维法把 n 维数据点降到 q 维数据点时采用的思想是保持数据集中数据点间的距离在两个不同维数空间中（近似）不变。投影寻踪试图找到数据集有令人感兴趣的分布的投影方向，在这些投影方向能展示出数据集的某些内在结构。该方法认为在数据集具有高斯分布的那些投影方向投影是最缺乏结构的，而非高斯分布所对应的投影方向能展示数据集结构[32]。

2.3 类数目的确定

一个数据集的数据点可以分为多少个类（子结构），一直是聚类分析的一个研究热点，至目前为止，还没有一个很好的办法可以保证获得准确的类数目，这是聚类分析中一个较为关键和困难的问题[26]。通常确定类数目的方法是：先提出衡量数据集分类结果好坏的评估指标 VIS（Validity indices），指标可能只有一个也可能有多个[27]，然后对于类数目 r 从最小值 $r_{\min}$（通常可设为 2）开始，到用户设定的最大类数目 $r_{\max}$ 结束进行循环，对这个过程中的每个给定的类数目 r，执行 k 次聚类算法。运行 k 次是因为聚类算法多含有参数，对参数取不同的参数值可获得不同聚类结果。然后以类数目 r 为横坐标，以对应于类数目 r 的不同参数值聚类结果中计算得到的 VIS 最优值作为纵坐标，把对应于从 $r_{\min}$ 到 $r_{\max}$ 的这些点依次相连得到一个 plot 图（折线或曲线），若此曲线关于类数目 r 并非单调曲线，则选择曲线 VIS 值最大值（或最小值）所对应的 r 值作为“正确”的类数目。若曲线单调，则选择曲线上局部地区 VIS 值有意义的突变点（称之为 knee 或 elbow）处所对应的类数目作为“正确”的类数目。对于这种“knee”现象（以 knee 作为选择类数目的根据），Tibshirani[24] 作出了理论上的解释，并由此提出了用 gap statistic（即 r 个类的“类内距离的平均值”之和的对数函数的负离差）的优化来估计正确的类数目这一方法。另外一种情形是算法的参数集中并没有类数目 r 这一参数，此时选择在参数集变化范围内始终保持类数目值不变的最大子参数范围对应的类数目作为正确的类数目。在类数目的确定过程中，有时往往需要计算多个不同定义的 VIS 值来综合考虑分析，以得出较合理的类数目。

在基于谱分析的聚类方法中，一些学者提出了用矩阵的扰动理论来自动获得类数目的方法[39]。指出当数据集中的 cluster 内部有较好的致密性而 cluster 之间有较好的分离性时（从定义的亲密度的角度来看），数据集的类数目等于亲密度矩阵大于 1 的特征值的个数。

在基于模型的聚类算法中，类数目的确定是通过“模型选择”来进行的。模型选择是在数据拟合精度与模型复杂性之间的折中，符合 Occam 剃刀原理：简单模型只在“有限范围”内做预测，而复杂模型能在更宽范围内做预测，但在“有限范围”内，复杂模型预测不如简单模型强。因此可结合两者的优点，通过简单模型的复合叠加得到相对复杂模型。通过选择不同的模型（包括类数 G 这个因素），对数据集进行分类，对不同模型的分类结果计算其 AIC 或 BIC 值[29]。然后以类数目为横轴，以不同模型的 AIC 值（或 BIC 值）作为纵轴，给出不同模型的 plot 图，选择 AIC 值或 BIC 值最优的那个模型及所对应的类数目作为数据集的“正确”类数。因此，目前类数主要是通过 VIS 的最大（小）值或 knee 点、AIC 值和 BIC 值的最大值、亲密度矩阵的特征值等来确定。多边形及网格剖分研究中的 PPOS 系统[1] 主要采用亲密度矩阵的特征值变换趋势来实现类数目的自动确定。

2.4 算法评估

算法评估[24-29] 讨论如何对聚类算法的优劣性作一个评价。由于聚类结果遵循的一个原则是“类内相似度尽可能大，而类间相似度尽量小”，因此很多对聚类算法的评估方法都是基于这一原则的。通常通过计算 VIS（Validity Indices）量化

值来衡量分类结果符合上述原则的程度，从而对算法作出优劣性的评估。VIS 可以分为 3 类[27]：外部准则(external criteria)、内部准则(internal criteria)和相对准则(relative criteria)。外部准则是把算法分类结果和(外部已知的)标准答案相比较，从而得出算法分类结果的正确性如何。该方法通常把 VIS 看成统计量，用 Monde Carlo 方法模拟出该统计量的分布曲线，然后用假设检验法对统计量的结果值(观测值)与显著性水平对应的临界值比大小，以确定现有分类结果和标准答案是否相似或一致。常见的外部准则 VIS 有：Rand 统计量、Huberts gama 统计量、标准 Huberts gama 统计量、Jaccard 系数和 FM 指标。内部准则以 VIS 对数据集内部量之间的比较来评判算法分类结果的好坏，例如用类信息矩阵 $\boldsymbol{C}$，元素 $C_{ij}=1$ 表数据点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 属同一类，而 $C_{ij}=0$ 表不同类。则 $\boldsymbol{C}$ 阵表示算法分类结果，若矩阵 $\boldsymbol{P}$ 是原始数据集的亲密度矩阵，则可定义合适的 $\boldsymbol{VIS}$ 来衡量矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{C}$ 的相似度，从而对分类结果作出优劣判断。常见的内部准则 VIS 有：CPCC、Huberts gama 统计量和其标准化。和外部准则一样，内部准则也常用假设检验法判断。而相对准则则对算法关于参数集在某个范围内进行循环执行，分为参数集中包含类数目参数和不包含类数目参数两种情况讨论，利用前面的“knee”等方法确定类数目，从而获得对应的一个分类结果，该结果被认为是参数集取不同值时相对最好的，然后用合适的 VIS 对结果评估。该法由于不需用假设检验，和上述两种方法相比，计算量要小很多。常见的 VIS 包括[26,27]：Dunn and Dunn-like 指标，RMSSDT、SPR、RS、CD 的组合，PC(Partition Coefficient)，PE(Partition Entropy Coefficent)，Xie-Beni index 等等。这 3 类准则中，外部准则最为客观，相对准则计算量最小，因此就我们的观点而言，外部准则优于相对准则，而相对准则要优于内部准则。在关于多边形有意义的剖分[1]的研究中，对我们剖分算法的评估采用了外部准则，标准答案是 80 多个多边形的人工剖分标准数据集[36]，采用统计学中的标准误差(standard error)来评估算法剖分结果和人工剖分结果的吻合程度，以评估算法及剖分结果的优劣。

结束语 本文结合我们在多边形与网格剖分中的研究[1]系统地对聚类算法研究中的 4 个关键问题进行了论述，包括聚类算法、距离函数、类数目的确定和算法评估等，基本涵盖了当前聚类研究的主要内容与主要研究方法。像谱分析、亲密度传播(AP)法、顺磁聚类法、基于分形的方法、复杂网络聚类法、仿生法和核聚类法都是很有特色的聚类算法，在各个学科中得到了广泛的应用，具有广阔的发展前景。聚类研究中的高维数、孤立点、噪声、数据信息缺失和 cluster 间的可覆盖性等增加了聚类亲密度定义和聚类算法获得正确结果的难度，且算法输入参数要尽量少。亲密度如何定义，类数目如何自动准确获得始终是聚类算法中最为困难的两个关键问题，我们在模式识别中的多边形与网格有意义的剖分研究中深刻地体会到了这两个问题的重要性。

当前，聚类算法的主要研究对象还是针对线性具有整数维数几何结构的数据集。而自然界中更普遍存在的是具有分数维数的几何结构，因此我们认为对具有分数维数的非线性几何结构的数据集聚类研究是未来一个重要的发展方向。

参 考 文 献

[1] Jin Jian-guo. PPOS SYSTEM: A System of Partitioning Polygonal Objects[C]// ICISE 2009. 2009: 920-923

[2] Wazavkar S V, Manjrekar A A. Text Clustering Using HFREC-CA and Rough K-Means Cluster Algorithm[J]. Discovery, 2014, 15(40): 44-47

[3] Zhang Chun-fei, Fang Zhi-yi. An Improved K-means Clustering Algorithm[J]. Journal of Information & Computational Science, 2013, 10(1): 193-199

[4] Zhong Luo, Tang Kun-hao, Li Lin, et al. An Improved Clustering Algorithm of Tunnel Monitoring Data for cloud Computing [J]. The Scientific World Journal, 2014

[5] Trikha P, Vijendra S. Fast Density Based Clustering Algorithm [J]. International Journal of Machine Learning and Computing, 2013, 3(1): 10-12

[6] Wu Jia-wei, Li Xiong-fei, Sun Tao, et al. A density-based clustering algorithm concerning neighborhood balace[J]. Journal of Computer Research and Development, 2010, 47(6): 1044-1052

[7] Frey B J, Dueck D. Clustering by Passing Messages Between Data Points[J]. Science, 2007, 315: 972-976

[8] Brusco M J, Kohn H-F. Comment on “Clustering by Passing Messages Between Data Points” [J]. Science, 2008, 319: 726

[9] Liu Rong, Zhang Hao. Segmentation of 3 D Meshes through Spectral Clustering[C]// Proceedings of the 12th Pacific Conference on Computer Graphics and Applictaions. 2004

[10] Blatt M, Wiseman S, Domany E. Superparamagetic Clustering of Data[J]. Physical Review Letters, 1996, 76(18):

[11] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases[C]// Proc. ACM SIGMOD Int. Conf. Management of Data. 1998: 73-84

[12] Guha S, Rastogi R, Shim K. ROCK: A Robust Clustering Algorithm for Categorical Attributes[C]// Proceedings of the IEEE Conference on Data Engineering. 1999

[13] Ng R, Han J. Efficient and Effictive Clustering Methods for Spatial Data Mining[C]// Proceeding's of the 20th VLDB Conference. Santiago, Chile, 1994

[14] Huang Z. A Fast Clustering Algorithm to Cluster very large Categorical Data Sets in Data Mining[C]// DMKD. 1997

[15] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases[C]// Proc. ACM SIGMOD Conf. Management of Data. 1996: 103-114

[16] Sharan R, Shamir R. CLICK: A clustering algorithm with applications to gene expression analysis[C]// Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology. 2000: 307-316

[17] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise” [C]// Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD'96). 1996: 226-231

[18] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60 (1), 208-221

[19] Wang W, Yang J, Muntz R. STING: A Statistical Information Grid Approach to Spatial Data Mining[C]// Proceedings of 23rd VLDB Conference. 1997: 186-195

[20] Kaufman L, Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis[M]. Wiley, 1990

[21] Duda R O, Hart P E, Stork D G. Pattern Classification, Second Edition[M]. A Wiley-Interscience Publication, 2001

[22] Peng Jing, et al. A new similarity computing method based on concept similarity in Chinese test processing[J]. Science in China Series F: Information Sciences, 2008, 51(9): 1215-1230

[23] Wang Hai-xun, Wang Wei, Yang Jiong, et al. Clustering by Pattern Similarity in Large Data Sets[C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. 2002: 394-405

[24] Tibshirani R, et al. Estimating the number of clusters in a data set via the gap statistic[J]. J. R. Statist. Soc. B, 63, part 2, 2001: 411-423

[25] Fraley C, Raftery A E. How Many Clusters? Which Clustering Method? Answer Via Model-Based Cluster Analysis[J]. The Computer Journal, 1998, 41(8): 578-588

[26] Sun Hao-jun, Wang Sheng-rui, Jiang Qing-shan. FCM-Based Model Selection Algorithms for Determining the Number of Clusters[J]. Pattern Recognition, 2004(37): 2027-2037

[27] Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques[J]. Journal of Intelligent Information Systems, 2001, 17(2/3): 107-145

[28] Fraley C, Raftery A E. Model-Based Clustering, Discriminant Analysis, and Density Estimation[J]. Journal of the American Statistical Association, 2002, 97(458): 611-631

[29] LEroux B G. Consistent Estimation of a Mixing Distribution [J]. The Annals of Statistics, 1992, 20(3): 1350-1360

[30] Baya A E, Granitto P M. ISOMAP based metrics for clustering [J]. Inteligencia Artificial, 2008, 12(37): 15-23

[31] Tenenbaum J B, de Silva V, Langford J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction[J]. Science, 2000, 290: 2319-2323

[32] Hyvarinen A, Oja E. Independent Component Analysis: algorithms and applications[J]. Neural Networks, 2000, 13: 411-430

[33] Alter O, Brown P O, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling[J]. PNAS, 2000, 97(18): 10101-10106

[34] Kim P M, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data[J]. Genome Res, 2003, 13(7): 1706-1718

[35] Murtagh F, Starck J L, Berry M W. Overcoming the Curse of Dimensionality by Means of the Wavelet Transform[J]. The Computer Journal, 2000, 43: 107-120

[36] De Winter J, Wagemans J. Segmentation of object outlines into parts: A large-scale intgrative study[J]. Cognition, 2006, 99: 275-325

[37] Freixenet J, Munoz X, Raba D, et al. Yet Another Survey on Image Segmentation: Region and Boundary Information Integration [C]// ECCV2002. LNS2352, 2002: 408-422

[38] 李瑞, 邱玉辉. 基于离散点的蚁群聚类算法的研究[J]. 计算机科学, 2005, 32(6): 111-113

[39] 田铮, 李小斌, 句彦伟. 谱聚类的扰动分析[J]. 中国科学 E 辑: 信息科学, 2007, 37(4): 527-543

[40] Barbara B, Chen Ping. Using the Fractal Dimension to Cluster Datasets[C]// Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge discovery and data mining(KDD-2000). ACM Press, 2000: 260-264

[41] 杨博, 刘大有, Liu Ji-ming, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66

[42] 张莉, 周伟达, 焦李成. 核聚类算法[J]. 计算机学报, 2002, 25(6): 587-590

[43] Al-Shammary D, Khalil I, Tari Z. A distributed aggregation and fast fractal clustering approach for SOAP traffic[J]. Journal of Network and Computer Applications, 2014, 41: 1-14

[44] Basu B, Srinivas V V. Regional flood frequency analysis using kernel-based fuzzy clustering approach[J]. Water Resources Research, 2014, 50(4): 3295-3316

[45] Li Xiang, Wong Hau-san, Wu Si. A fuzzy minimax clustering model and its applications[J]. Information Sciences, 2012, 186(1): 114-125

(上接第 283 页)

结束语 本文对给定整数的最佳 BSD 表示的性质进行了深入研究，给出了计算给定整数的随机最佳 BSD 表示的算法，所得算法能快速产生给定整数的随机最佳 BSD 表示。后续工作中，将利用本文的结论设计抗边信道攻击的 ECC 快速标量乘法算法。

参 考 文 献

[1] Ebeid N, Hasan M A. On binary signed digit repress-entations of integers[C]// Design Code Cryptogr. 2007, 42: 43-65

[2] 李忠, 彭代渊. 整数的带符号二进制表示数的快速计算[J]. 计算机应用, 2012, 32(11): 3121-3124

[3] Wu T, Zhang M, Du H, et al. On optimal binary signed digit representations of integers[J]. Applied Mathematics, 2010, 25(3): 331-340

[4] Ganesan P, Manku G S. Optimal routing in Chord[C]// Proc. 15th ACM-SIAM Symposium on Discrete Algorithms (SODA 2004). 2004: 169-178

[5] Sawada J. A Gray code for binary subtraction[C]// 2nd Brazilian Symposium on Graphs, Algorithms and Combinatorics(GRACO 2005). 2005

[6] Manku G S, Sawada J. A Loopless Gray Code for Minimal Signed-Binary Representations[C]// Brodal G S, Leonardi S, eds. ESA 2005. LNCS 3669, 2005: 438-447

[7] Hankerson D, Menezes A, Vanstone S. Guide to elliptic curve cryptography[M]. Springer-Verlag Professional Computing Series, 2004

[8] Avanzi R M. A note on the signed sliding window integer recoding and a left-to-right analogue[C]// Handschuh H, Hasan A, eds. SAC 2004. LNCS 3357, 2004: 130-143

[9] Joye M, Yen S M. Optimal left-to-right binary signed-digit recoding[J]. IEEE Transactions on Computers, 2000, 49: 740-748

[10] Okeya K, Schmidt-samoak C, Spahn, et al. Signed binary representations revisited[C]// Andances in Cryptology-CRYPTO'04. LNCS 3152, 2004: 123-139