统计聚类模型研究综述

管 涛
(郑州航空工业管理学院计算机科学与应用系 郑州 450015)

摘 要 聚类分析在工程领域如生物序列分析、图像分割、文本分析等广泛应用。聚类方法涉及广泛,而基于概率统计理论的方法是其中的一大类。从最基本的FCM模型出发,阐述了势函数(Potential)、山脉(Mountain)函数聚类方法、信息熵方法,分析比较了这些方法的适用范围和优缺点,介绍了当今流行的核聚类、谱聚类和高斯混合模型聚类方法及其求解过程,并分析了它们的优缺点、计算复杂性等指标。最后,介绍了一些崭新的聚类模型的研究方向。

关键词 聚类分析,统计学习,高斯混合模型,谱聚类,核聚类
中图法分类号 TP301 **文献标识码** A

Overview of Statistical Clustering Models

GUAN Tao
(Department of Computer Science and Application, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450015, China)

Abstract Clustering analysis is widely applied to engineering fields, such as biology sequence analysis, image segmentation, text analysis. Currently there have been many clustering methods and statistical learning based methods constitute a class of them. This paper started from FCM, introduced classical methods, such as potential and mountain functions, entropy method, and then analyzed their properties and applicability. Moreover, we also introduced the state-of-art clustering techniques, such as kernel clustering, spectral clustering and Gaussian mixture model based clustering, narrated the solving process and analyzed their properties, computation complexity. At last, this paper presented several research directions.

Keywords Clustering analysis, Statistical machine learning, Gaussian mixture models, Spectral clustering, Kernel clustering

1 引言

近几十年来,随着信息技术的快速发展,各个工程领域产生了大量的数据,如网络上的网页信息、医疗检查中的超声图像、军事领域的各种雷达信号、卫星及飞机上获取的遥感图像、潜艇的声纳信号等。这些数据具有海量、实时、高维、分布未知、潜在相关等特点,因而需要发展崭新的非参数数据分析方法。

聚类是一个非参数数据分类方法,不需要先验信息,在自动控制、模式识别、信息处理中广泛应用。它使用某种度量将数据分为有限的类,这些度量包括线性与非线性距离、熵、包含度等。针对不同的研究对象,聚类模型和算法具有不同的形式,同时在时间和空间效率上也有较大差别[1,2]。

除经典的聚类模型外,近些年来,涌现了不少崭新的聚类方法,如谱聚类、空间约束高斯混合模型、核聚类方法、高斯过程等。在理论上,这些模型具有深入的数学基础,同时也具有广泛的应用范围,如图像配准、分类,文本聚类等。文中将对这些新型的聚类做详细的介绍。

2 经典聚类模型

2.1 模糊k均值来源及模型

首先引述模糊聚类的数学背景,给出文献[3]中表述的聚类定义。假定类别数目 k 已给定,则聚类问题最直接的表示方式为:给定 n 个数据 $\{x_i\}_{i=1}^n \subset R^n$ 以及类别数 k,优化如下的问题:

$$(P_1)\quad \min_{c_l,u_{il}} \sum_{i=1}^{n}\sum_{l=1}^{k}(u_{il})^m \| x_i-c_l \|_2^2 \qquad (1)$$

s. t.

$$\sum_{l=1}^{k}u_{il}=1, u_{il}\geqslant 0, i=1,\cdots,n, l=1,\cdots,k$$

该算法被称为模糊 k-均值算法(或称FCM算法)。而C-均值算法是 $u_{il}\in\{0,1\}$ 时的情形。

在FCM,每一个数据点满足基本的概率关系约束,即 $e^Tu_j=1$,其中 $e=(1,1,\cdots,1)^T$,可知模糊变量 u_{ij} 位于一系列的超平面上。模糊 k 均值计算模型如(P1)、目标函数(1)构成

到稿日期:2011-10-11 返修日期:2011-12-25 本文受国家自然科学基金项目(41171341),教育部新世纪优秀人才支持计划,河南省科技创新杰出青年计划(114100510006),航空科学基金光电控制技术国防科技重点实验室资助项目(20095155008),河南省科技厅科技攻关项目(122102210227),河南省科技厅基础与前沿技术研究计划项目(092300410140),河南省教育厅项目(2011B520038,2010B520032),郑州市科技局项目(112PPTGY248-6)资助。

管 涛(1974-),男,博士,讲师,CCF会员,主要研究方向为统计机器学习、数据挖掘、图像处理,E-mail:guantao@tsinghua.org.cn。

了一个平方误差评价函数。其中，u_{ij} 表示数据点 x_j 对于聚类中心 c_i 的隶属程度，它们构成了 $\mathbf{X}$ 的一个模糊分割矩阵 $U=(u_{ij})_{c\times n}$。模糊系数 $m\in(1,\infty)$，$d^2(x_j,c_i)$ 表示特征点 x_j 到中心 c_i 的距离。在 FCM 模型中目标函数为双变量的，固定一个变量，如 u_{ij}，则得到的是一个凸规划问题，Selim 和 Ismail 等解释了目标函数的凸性[4]。一般的求解方法是：利用 Lagrnge 函数简化目标模型，然后利用一阶最优化理论的 K-T 条件得到 u_{ij} 的计算表达式：

$$u_{ij}=\left(\sum_{l=1}^{k}\left(\frac{d(x_j,c_i)}{d(x_j,c_l)}\right)^{\frac{2}{m-1}}\right)^{-1}=\left(\sum_{l=1}^{k}\left(\frac{d_{ij}}{d_{lj}}\right)^{\frac{2}{m-1}}\right)^{-1} \tag{2}$$

和

$$c_i=\frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m} \tag{3}$$

当聚类中心处于某一个数据点时，式(2)中分母的距离可能为 0，在分析算法的时候，这可以看作一个极限情况。如果在目标函数中用 $e^{d_{ij}}$ 代替 d_{ij}，就不存在这种情况。FCM 的时间复杂度为 $O(knl)$，其中：k、n、l 分别为聚类的数目、数据点数和迭代次数。

模糊 k 均值(FCM)来源于数据的统计分析，评价函数描述了数据点分类的误差总和，在概率约束条件下使得分类到达最优，是最常用的一种聚类方法，具有较快的收敛速度和一定的健壮性，比较适合于大规模数据集合信息处理。但是它容易受噪声影响和陷入局部极小点或鞍点。文献[5]讨论了 FCM 的局部收敛性和极值点特性问题。

FCM 模型在符号数据(categorical data)上的扩展称为 k-modes 算法[6]，它使用了符号匹配计算对象之间的不相似性和模式(patterns or modes)，建立了大规模符号数据集聚类模型。还有其他一些 FCM 扩展模型。根据模糊成员关系和概率典型性，Pal 等给出了 FPCM 算法[7]，它是一种混合模型，属于 FCM 的一种变化形式，在噪声环境下改善了聚类的效果，但是并没有改变 FCM 的本质缺点。Belacel 等人给出了模糊 J-means 算法，它提高了 FCM 的计算效率[8]。Chepoi 等给出了在数据集上存在结构约束的模糊聚类模型[9]。Likas 等人给出了动态环境下增量式聚类的全局 FCM 模型与算法[10]。这些方法在某些方面提高了聚类结果的质量，但避免不了 FCM 方法本身的一些缺陷，如先验参数的选择。

2.2 势(Potential)函数聚类方法

物理学研究表明空间粒子本身存在势能并与其他粒子发生作用，这种能量通过势函数得以表达。借鉴这个基本原理，如今，势函数在模式识别、机器学习领域中有着较为广泛的应用，如自动模式分类、模拟退火算法、聚类等。在聚类应用中，若将空间中的数据点或网格点看作能源，则在密度大的区域(数据稠密的区域)能量值就高，反之，其能量值就低。典型的势函数[11]如：$P(x,x_k)=\exp\{-\alpha\|(x-x_k)\|^2\}$ 和 $P(x,x_k)=\frac{1}{[1+\alpha\|(x-x_k)\|^2]}$。利用 Potential 函数来定义模糊聚类的评价函数，如 $P_t(B,X)=\sum_{k=1}^{n}\exp\{-\alpha\|(x-x_k)\|^2\}$。通过距离公式，势函数 $P_t(B,X)$ 给出了点 x 处的总能量。

Mountain 方法是一个近似聚类方法，通过势函数表示数据点分布密度[12]。Mountain 方法中的对象集合 $X=\{x_1,x_2,\cdots,x_n\}$ 被一系列网格划分，在每个格点 N_i 上计算 Mountain 函数值(α 为常数)：

$$M(N_i)=\sum_{k=1}^{n}e^{-\alpha d(x_k,N_i)}$$

Mountain 函数值表明了该格点附近数据点的密度。函数值越大，则密度值越高。首次达到最大值的点即是第一个近似聚类中心，即求：$M_1^*=\max_i M(N_i)$，然后利用下式迭代计算其他近似聚类中心(β 为常数，k 为迭代步)：$M^k(N_i)=M^{k-1}(N_i)-M_{k-1}^*\sum_{k=1}^{m}e^{-\beta d(N_{k-1}^*,N_i)}$。Mountain 方法的计算复杂性为：$O(mn+cm)$，其中 m,n,c 分别为格点数、数据点数和聚类数。其次，Mountain 方法对人为参数 α,β 较为敏感。α 较大，易产生单个尖峰；α 较小，易产生许多小的尖锋。如今，这种势函数被推广，一些作者直接在数据点上求势函数值[12]，这种方法的计算复杂性为：$O(n^2+cn)$。势函数方法的优点是不需要事先指定聚类数目，但计算复杂性比 FCM 高。

2.3 熵函数聚类方法

近年来，信息理论聚类方法成为新的研究方向，一些应用包括文本挖掘、视频序列匹配、图像分段等。作为一个非距离度量，信息熵在模糊聚类中应用广泛[13-16]。基于确定性退火(deterministic annealing)方法，Beni 和 Liu 给出了最小偏差模糊聚类方法(LBFC)[16]，聚类熵函数表示为：

$$S=-\sum_{j=1}^{n}p_j(x_j,c_i)\log p_j(x_j,c_i)$$

式中，$p_j(x_j,c_i)$ 表示中心 c_i 包含 x_j 的概率。在约束关系 $\sum_{j=1}^{n}(x_j-c_i)p_j(x_j,c_i)=0$ 下，最小化该函数求得概率值，聚类中心表达为：$c_i=\sum_{j=1}^{n}p_j(x_j,c_i)x_j$。在某种意义上，LBFC 与 PCM 的迭代计算公式是等价的[17]，且 LBFC 容易产生重叠聚类。Yao 等人提出了基于 Shannon 熵度量的模糊聚类模型[13]。该模型计算在每个数据点的熵值，优点是不需要预先指定聚类数目。数据点 x_i 处的熵值为：

$$E_i=-\sum_{j\in X,j\neq i}(S_{ij}\log_2 S_{ij}+(1-S_{ij})\log_2(1-S_{ij}))$$

式中，$S_{ij}=e^{-\alpha D_{ij}}$，$D_{ij}$ 是数据点 x_i 和 x_j 之间的距离。

熵 E_i 也可看作 x_i 点处的势能，因此，除了函数定义的差别外，它与在数据点上网格划分的 Mountain 方法的基本思想一致。这个方法不需要预先指定聚类数目 k，更适合聚类小规模静态数据。在动态环境下，它无法利用前次计算结果等先验信息，需要重新计算每个点的熵值，具有较高的计算复杂性。张志华等人根据 FCM 算法，使用极大熵原理重新构造了聚类算法的迭代公式[18]。概括来讲，它们是 ACE 聚类方法的特殊形式[20]。为实现数据集的非线性分割，Gokcay 等人使用信息度量来估计分割数据集，提出一个新颖的峡谷搜寻聚类算法[19]，其中的评价标准来自 Renyi 熵估计器。与经典方法寻找数据最稠密的区域不同，它寻找数据分布稀少的区域(即峡谷)形成数据的非线性分割，提高数据集的分割精度，但是需要预先指定聚类数目 k，且算法具有较高的计算复杂性。常选的非线性函数为 Guass 核函数。还有一些其他的改进 FCM 的方法，如利用熵方法改进评价函数，Karayiannis 提出了最大熵聚类算法 (MECA)[14]；利用最大熵重新定义评价函数，Li 等人给出了一种熵聚类方法[15]等。

3 核聚类模型

核方法用核函数实现，用来将数据从低维特征空间映射到高维特征空间，然后，在高维空间中进行各种运算，如聚类、

成分分析[21]、判别分析[22]。核聚类方法是模式识别领域流行的算法[23,24]，在视频分析、医疗图像分割、文本分析、异常检测中得到广泛的应用。典型的核聚类方法包括核 k-means 算法、Merce 核聚类[24]、支撑向量机、多核聚类方法[25,26]等等。

从原理上讲，核方法通过非线性映射 ϕ 将数据点从低维空间映射到合适的高维空间，然后在高维空间实现数据的线性划分。图 1 演示了数据映射的过程。原始训练集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 变为 $\{(\phi(x_1),y_1),(\phi(x_2),y_2),\cdots,(\phi(x_n),y_n)\}$。例 1 给出了一个非线性映射过程。

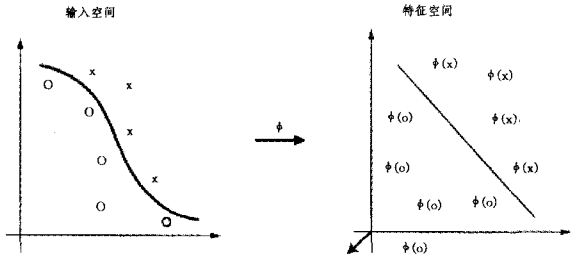

图 1 输入数据经过非线性映射 ϕ 变换到高维特征空间，实现线性可分

例 1 给定如下非线性映射[27]：二维到三维映射：$(x_1,x_2)\xrightarrow{\phi}(x_1^2,x_2^2,\sqrt{2}x_1x_2)$，该映射将数据样本从二维坐标映射到三维坐标。如类标签为 1 的训练样本 $((4,3),1)$ 映射为 $((16,9,17),1)$。

在实际应用中，ϕ 是未知的，那么，如何在特征空间中计算数据之间的相似性呢？向量相似性可以通过内积来定义，但是，特征空间的向量值是未知的。核函数提供了一种有效的解决方法，特征空间向量的内积可以通过核函数表达，即 $\langle\phi(x),\phi(y)\rangle=k(x,y)$。在经典的支撑向量机模型中隐含地采用了这种方法进行分类计算。

3.1 核 k 均值

核 k 均值算法是对经过核函数 ϕ 映射后的数据使用 k-means 算法聚类，在一些文献中有所论述[28,29]。这里简要阐述核 k-means 算法的基本原理。给定空间 $\mathbf{R}^n$ 中的数据 $X=\{x_1,x_2,\cdots,x_n\}$ 和非线性映射 ϕ，经过映射后的数据变为 $\phi(X)=\{\phi(x_1),\phi(x_2),\cdots,\phi(x_n)\}$，在核空间中聚类 $\phi(X)$ 即可但是，通常 ϕ 是未知的，通过核函数 $K(\cdot,\cdot)$ 表达，即 $K(x,y)=\langle\phi(x),\phi(y)\rangle$。类似于普通的 k-means 算法，考虑原始空间中的 k 个类别 $C_1,C_2,\cdots,C_k$ 及其在核空间的类中心 $c_1,c_2,\cdots,c_k$，最小化目标函数

$$\min\sum_{x_i\in C_j}\sum_{j=1}^{k}\|\phi(x_i)-c_j\|_2^2$$

式中，$c_j=\sum_{\phi(x_j)\in C_j}(\phi(x_j)/(C_j))$，$|\cdot|$ 表示集合元素的数目。可以将目标函数转化为核函数形式[28]，即：

$$\begin{aligned}\|\phi(x_i)-c_j\|_2^2&=\langle\phi(x_i)-\phi(c_j),\phi(x_i)-\phi(c_j)\rangle\\&=\langle\phi(x_i),\phi(x_i)\rangle+\langle c_j,c_j\rangle-2\langle\phi(x_i),c_j\rangle\\&=\langle\phi(x_i),\phi(x_i)\rangle+\sum_{x_i,x_j\in C_j}\langle\phi(x_i),\phi(x_j)\rangle/|C_j|^2-2\sum_{x_j\in C_j}\langle\phi(x_i),\phi(x_j)\rangle/|C_j|\\&=K(x_i,x_i)+\sum_{x_i,x_j\in C_j}K(x_i,x_j)/|C_j|^2-2\sum_{x_j\in C_j}K(x_i,x_j)/|C_j|\end{aligned}$$

对于初始随机划分的聚类和权重，构造迭代公式：$c^*(x_i)=\arg\min_c\|\phi(x_i)-c\|$，$C_j=\{x_i:c^*(x_i)=c_j\}$，迭代计算即可得到最优聚类。Dhillon 等人引入了权重的核 k-means 算法，聚类中心表示为样本的权重和，类似于模糊 k-means 算法中的隶属度。进而，论述了权重的核 k-means 与谱聚类之间存在等价关系，即在矩阵的迹所定义的目标函数下，这两个模型的目标函数是等价的。Filippone 等人总结了近些年来的一些核聚类方法，阐明了核聚类与谱聚类之间的关系。

3.2 多核学习模型

常见的分类算法采用了单个核函数，多个单核的加权组合构成了当前流行的多核学习方法(MKL)[25,26,31-34]。有学者指出[25]，相比单核学习，多核学习可以提高分类器的性能。多核学习常用于支撑向量机模型中。给定数据点 $X=\{x_1,x_2,\cdots,x_n\}$，非线性最优划分函数可以表示为：$f(x)=\sum_{x_i\in SV}\alpha_i^*K(x_i,x)+b^*$，其中：$SV$ 是支撑向量集，α_i^*，b^* 为最优的系数。在多核学习中，单个核函数变为多个核函数 $K_m(\cdot,\cdot)$ 的凸组合形式，即：

$$K(x,x')=\sum_{m=1}^{M}\gamma_mK_m(x,x')$$

式中，$0\leqslant\gamma_m\leqslant1$，$\sum_{m=1}^{M}\gamma_m=1$。每个核函数 $K_m(x,x')$ 存在于一个再生核 Hilbert 空间(RKHS) $\mathcal{H}_m$ 中，而 $K(x,x')$ 定义于各子空间的直和中，即：$\bigoplus_{1,\cdots,M}\mathcal{H}_m$，当取多核核函数时，支撑向量机模型的优化目标变为[26]：

$$\min\sum_m\frac{1}{\gamma_m}\|f_m\|_{\mathcal{H}_m}^2+C\sum_i\xi_i$$

s. t.

$$y_i(\sum_m f_m(x_i)+b)\geqslant1-\xi_i,\xi_i\geqslant0,\sum_m\gamma_m=1,\gamma_m\geqslant0$$

权重系数 γ_m 采用 l_1 范数保证了多核的权重是稀疏的。该目标函数的求解涉及较为深入的优化方法，可以参看 Rakotomamonjy 等人的论文。多核支撑向量机模型可用于聚类分析、密度估计、图像分类、视频分析等领域。

4 谱聚类算法

谱聚类在不规则分布数据聚类方面的优良效果吸引了众多学者的注意，成为当今机器学习领域研究的热点之一[35]。谱聚类以图论、矩阵分析为基础，研究相似矩阵的拉普拉斯矩阵的特征值和特征向量，构建高维空间数据的低维表示，在图像分割、视频分析等应用领域效果显著。

国内外在这方面不断有崭新的研究成果出现。美国加州大学 Jordan 的研究组从图模型的角度研究谱聚类，在理论分析和应用方面取得了一系列的成果。A. Y. Ng 等人给出了谱聚类的理论分析，证明了拉普拉斯矩阵的前 k 个特征向量构成的矩阵的行向量空间以 k 个正交的向量为中心[36]。Y. Weiss 简单地总结了近些年来几个比较好的谱聚类算法，给出了实验比较分析[37]。在谱聚类的随机分析方面有一些崭新的研究工作，如 Meila 和 Shi 的关于谱聚类的随机游走分析工作[38]。谱聚类和图分割有紧密联系，一个关键之处在于计算相似矩阵的特征值和特征向量，在图像分割领域，有一些著名分割算法，如 Normailized Cut(NCUT)、ratio-cut、比率关联标准(ratio association criteria)。但是，当图的节点较多时，计算特征值和特征向量是很耗时、耗内存的。最近几年，为了解决经典谱聚类的处理速度和超大图的分割问题，出现了一些多层谱聚类方法，并在图像分割中取得了很好的效果。Kong 等人研究了谱聚类的时间效率问题，提出了多层的谱聚类算法，并用于图像分割[39]。Dhillon 等人提出的算法直接优化

图分割目标函数，避免了耗时的特征值的计算[28]。国内的一些学者在谱聚类领域也取得了较好的成果。西安电子科技大学焦李成教授的研究组在谱聚类图像分割方面有不少工作，如基于免疫谱聚类的图像分割[40]、多尺度谱聚类、空间一致性约束谱聚类算法等。西北工业大学的田铮教授等人先后在谱聚类的算法和应用方面取得了较突出的成果[41]，利用向量之间以内积度量的相似性，提出了多尺度谱聚类算法。目前，谱聚类方法有几个典型的问题仍值得继续研究：多尺度相似性度量的设计；由于谱聚类具有 $O(n^3)$ 的计算复杂度，因此，它在海量数据集上运行效率低，需要设计新的算法；现有的多尺度谱聚类算法缺少严格的数学论证；抽样技术对聚类结果的影响缺少理论分析。

标准的谱聚类算法对每对样本点采用了固定方差 σ 计算相似性矩阵，这种情况下算法不能区分不同尺度的数据聚集。针对这个弱点，国内外学者提出了不同的解决方法。Zelnik-Manor 等人提出了自调节谱聚类算法[42]，该方法利用向量的第 K 个最近邻计算局部尺度，得到对象间不同尺度的相似性度量。

$$A_{ij}=\exp\{-\|x_i-x_j\|^2/\sigma_i\sigma_j\}$$

式中，$\sigma_i=\|x_i-x_K\|$，x_K 是 x_i 的第 K 个最近邻。

该算法中第 K 个样本点的选择具有较大的随机性，聚类结果变化较大。李小斌等人提出了一种层次谱聚类算法，尺度选择采用了从粗到细的简单方式[43]。该算法通过采样定理保证了每个类别得到样本的最低采样量，但是，没有讨论和分析抽样方法对聚类结果的影响。

Bengio 等人证明了谱聚类与核 PCA 算法是等价的，且二者是核函数的特征函数学习问题的特例[44]。给定 Hilbert 空间中的特征系统 $K_q f_k=\lambda_k f_k$，它在离散样本情况下具有如下的形式：

$$\frac{1}{n}\sum_{i=1}^{n}K(x_j,x_i)f_k(x_i)=\lambda_k f_k(x_j)$$

式中，算子 K_q 定义为 $(K_q f)(x)=\frac{1}{n}\sum_{i=1}^{n}K(x,x_i)f(x_i)$，$K(x,x_i)$ 是一个对称核函数。

该特征系统也称为 Nyström 方法。Fowlkes 等人使用该特征系统在样本上计算压缩后的点对相似性，从而大大降低了谱聚类算法的时间复杂度[45]。在标准化的[0,1]区间，他们得到的 Nyström 展开为：

$$f_k(x_j)=\frac{1}{n\lambda_k}\sum_{i=1}^{n}K(x_j,x_i)f_k(x_i)$$

式中，f_k 是第 k 个特征函数，也可以简化表示为第 k 个特征向量。

Fowlkes 等人指出，该展开可以利用核函数插值将有限个样本点得到的特征向量扩展到任意点 x_j。对于特征向量而言，可以利用低阶矩阵的特征向量近似高阶矩阵的特征向量，来降低计算复杂性。假设相似矩阵 W 分解为：

$$W=\begin{bmatrix}A & B\\ B^T & C\end{bmatrix}$$

$A=U\Lambda U^T$，则 W 的近似特征向量为 $[U \quad B^TU\Lambda^{-1}]^T$，矩阵分量 $B^TU\Lambda^{-1}$ 即为向量形式的 Nyström 扩展。

5 高斯混合模型与 EM 算法

高斯混合模型是一种常用的统计模型，它由一系列的高斯分布加权平均构成，各个分支分布参数相互独立。常用的参数估计方法为极大似然估计，而计算该估计的方法包括 EM 算法、牛顿迭代法、交叉信息熵等。近些年来，高斯混合模型得到了广泛的关注，在数值逼近、语音识别、图像分类、图像降噪、图像重构、故障诊断、目标检测和跟踪、邮件过滤器等邻域得到应用，产生了良好的效果。下面介绍 GMM 的推理过程和两个应用：基于空间 GMM 模型的图像分段方法、点集漂移图像配准方法。

5.1 基本模型及其求解

混合高斯模型具有如下的形式：

$$f(x)=\sum_{i=1}^{k}\alpha_i G_i(x|\theta_i) \tag{4}$$

式中，$G_i(x|\theta_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{\|x-\mu_i\|^2}{2\sigma_i^2}\right\}$，$\theta_i=(\mu_i,\sigma_i)$；$\sum_{i=1}^{k}\alpha_i=1$，$\alpha_i>0$。

在实际应用中，通常假设数据符合混合高斯分布，然后，利用极大似然估计等方法计算混合高斯分布中的参数 θ_i 和 α_i。但是，直接对式(3)使用极大估计方法得到如下的计算公式：

$$\begin{aligned}l(\Theta)&=-\ln[\prod_{j=1}^{N}f(x_j)]=-\sum_{j=1}^{N}\ln f(x_j)\\&=-\sum_{j=1}^{N}\ln\sum_{i=1}^{k}\alpha_i G_i(x_j|\theta_i)\end{aligned} \tag{5}$$

最小化式(5)计算模型的参数值。通常使用 EM 算法进行计算，下面利用一阶导数导出参数 θ_i 和 α_i 的迭代公式。将 $l(\Theta)$ 分别对 μ_i，σ_i 求导并令其等于 0，得到：

$$\frac{\partial l(\Theta)}{\partial\mu_i}=\sum_{j=1}^{N}\frac{\alpha_i G_i(x_j,\theta_i)}{f(x)}\cdot\frac{(x_j-\mu_i)}{\sigma_i^2}=0$$

$$\frac{\partial l(\Theta)}{\partial\sigma_i}=\sum_{j=1}^{N}\frac{\alpha_i G_i(x_j,\theta_i)}{f(x)}\left\{\frac{1}{\sigma_i}-\frac{\|x_j-\mu_i\|^2}{\sigma_i^3}\right\}=0$$

如果假定 $\alpha_i=P(\theta_i)$，根据贝叶斯规则得到：

$$P(\theta_i|x_j)=\frac{\alpha_i G_i(x_j|\theta_i)}{\sum_{l}^{k}\alpha_l G_l(x_j|\theta_l)}=\frac{\alpha_i G_i(x_j|\theta_i)}{f(x_j)}$$，从而得到

$$\mu_i=\sum_{j=1}^{N}\frac{P(\theta_i|x_j)x_j}{P(\theta_i|x_j)}$$

$$\sigma_i^2=\frac{\sum_{j=1}^{N}P(\theta_i|x_j)\|x_j-\mu_i\|^2}{\sum_{j=1}^{N}P(\theta_i|x_j)}$$

在式(4)中，直接求解参数 α_i 的最优解不太容易。在现有的文献中，通常将 α_i 看作参数的先验概率，其求解方法不同。J. A. Bilmes 在技术报告中假定每个样本来自于某个单独的高斯分布简化模型，然后利用贝叶斯公式推导出 α_i 的计算公式[46]。Bishop 在《Neural Networks for Pattern Recognition》给出了另外一种推导方法，得到了同样的计算公式[47]。该方法利用 Jessen 不等式来构造目标函数的上界。最小化该上界同样可以达到计算最优参数的目的。这里直接引出 α_i 计算公式，有兴趣的读者可以参考以上文献。

$$\alpha_i=\frac{1}{N}\sum_{j=1}^{N}P(\theta_i|x_j)$$

5.2 GMM 在图像处理中应用

高斯混合模型在图像处理领域应用广泛，包括图像分割、图像配准、图像增强等。传统的高斯混合模型图像分割方法假设像素在特征空间中的分布满足某个高斯混合分布，然后，通过特征空间的点计算模型中的参数。通常使用 EM 算法求解最优参数，如上节所述。目前，混合模型在图像分割中有崭新的发展，主要侧重于空间约束条件下的混合模型及其在图像分割中的应用、广义 EM 算法、非高斯混合模型及其在图像

分割中应用、混合模型和回归模型的联系与应用。S. Sanjay-Gopal 等人将混合模型和马尔科夫随机场(MRF)相结合，构造了具有空间约束的像素级的图像分割模型[48]，并给出了扩展的 EM 算法。随机场模型考虑了像素之间的近邻关系，使得分支分布的比例参数(也被称为先验)发生变化。K. Blekas 等人改进了这个模型的约束函数，并利用约束优化算法提高了模型的求解速度[49]。随后，一些作者提出了新的改进措施。C. Nikou 等人在混合模型中分别融合高斯-马尔科夫先验和狄利克莱混合多项式先验来解决图像分割中的先验预分配问题[50,51]。由于空间约束下的混合模型在图像分割中运算速度可能较慢，因此，T. M. Nguyen 等人利用 Jessen 不等式设计了一种新的迭代算法来求解混合模型[52]。以上算法放弃了像素的独立性假设，考虑了像素与其 4 个近邻的约束关系。但是，主要的缺点包括：仅面向像素级别图像分割，领域结构简单，约束尺度单一，运行效率不高等。在混合模型的扩展方面，马江洪等人研究了指数族分布情况下的混合模型，给出了其信息几何性质，并应用于图像中线状模式的识别[53]。而指数分布族概括了高斯分布、指数分布、泊松分布等具体形式。向日华等人将高斯混合模型用于距离图像分割，产生了较好的效果[54]。在混合模型参数求解方面，比较多的文献研究了 EM 算法的各种扩展形式，如随机 EM 算法。下面引入 Sanjay-Gopal 等人的空间约束高斯混合模型 SVMM[48]，这个空间约束混合模型是其他空间模型的基础。

在一般的高斯混合模型中，系数 α_i 表示某个高斯分支的权重。在 SVMM 中，α_i 被细化到每个像素，即第 j 个像素 x_j 在第 i 个分支的权重为 α_i^j，从而得到如下的混合模型：

$$f(x_j)=\sum_{i=1}^{k}\alpha_i^j G_i(x_j|\theta_i),$$

SVMM 模型的求解采用了广义 EM 算法，定义了如下的 Q 函数：

$$Q(\Theta|\Theta^{(k)})=E\{\ln[\prod_{j=1}^{N}\prod_{i=1}^{M}[\alpha_i^j G_i(x_j|\theta_i)]]\}+\ln[f(\alpha_1,\cdots,\alpha_2,\Theta)]$$

式中，$f(\alpha_1,\cdots,\alpha_2,\Theta)$ 为参数的先验信息，定义为

$$f(\alpha_1,\cdots,\alpha_M,\Theta)=\exp[-U(\alpha^1,\cdots,\alpha^N)]/K_\beta$$

式中，$U(\alpha^1,\cdots,\alpha^N)=\beta\sum_{c\in C}V_c(\alpha^1,\cdots,\alpha^N)$，$C$ 表示图中所有的团(在 Markov 随机场中，团的定义有多种形式，邻近的两个像素组成的团是比较简单的形式)，$V_c(\cdot)$ 表示团势函数。$f(\cdot)$ 简化的形式为仅仅考虑相邻的像素有作用关系，定义如下：

$$f(\alpha_1,\cdots,\alpha_M,\Theta)=\frac{1}{Z_\beta}\exp[-\beta\sum_{(j,m)\in G}\sum_{i=1}^{L}(\alpha_i^j-\alpha_i^m)^2]f(\mu)f(\sigma)$$

式中，$\mu=(\mu_1,\cdots,\mu_M)$，$\sigma=(\sigma_1,\cdots,\sigma_M)$。$\alpha_i^j$ 满足约束条件 $0\leqslant\alpha_i^j\leqslant 1$，$\sum_{i=1}^{m}\alpha_i^j=1$。

最大化 Q 函数需要利用一些优化的求解方法。除了考虑每个像素点之外，得到的迭代公式类似于标准高斯混合模型的公式。

高斯混合模型可用于(3D)图像配准。近些年来，出现不少崭新的文献，其中基于高斯混合模型的点集配准算法比较著名[55]。下面简单介绍相关点飘移(Coherent Point Drift, CPD)算法如何运用高斯混合模型。对于需要配准的 D 维的点集 $\mathbf{X}=\{x_1,x_2,\cdots,x_n\}$ 和 $\mathbf{Y}=\{y_1,y_2,\cdots,y_m\}$，CPD 假设 $\mathbf{Y}$ 为高斯混合模型的中心，$\mathbf{X}$ 由这个混合模型产生，那么，在不考虑噪声的情况下，可以定义如下的混合模型：

$$f(x)=\sum_{i=1}^{m}\alpha_i p(x|\varphi_i)$$

式中，$p(x|\varphi_i)=\frac{1}{(2\pi\sigma^2)^{D/2}}\exp\left\{-\frac{\|x-y_i\|}{2\sigma^2}\right\}$，$\varphi_i=(y_i,\sigma)$。

在刚体点集配准中，y_i 被重新参数化为变换 $\mathbf{T}(y_m;\mathbf{R},\mathbf{t},s)=s\mathbf{R}y_m+\mathbf{t}$，其中：$s,\mathbf{R},\mathbf{t}$ 是参数。在仿射点集配准中，变换函数为 $\mathbf{T}(y_m;\mathbf{B},\mathbf{t})=\mathbf{B}y_m+t$，其中：$\mathbf{B},\mathbf{t}$ 为参数。在非刚体配准中，变换函数定义为 $\mathbf{T}(Y,v)=\mathbf{Y}+v(\mathbf{Y})$，其中：$v$ 为位移函数。在文献[51]中，关于这 3 种变换均有对应的参数计算方法，有兴趣可以深入阅读。

6 算法性能的分析与比较

以上部分讨论了一些典型的统计聚类模型和算法，表 1 给出了这些领域中一般方法的性能比较和优缺点、几个典型应用领域，其中复杂性主要从时间角度考虑。假设 n 为数据量，m 为数据空间网格划分的格点数，r 为抽样的样本数，k 为聚类数，l 为迭代步数，$\varepsilon>0$，$D=\sum_{k=1}^{M}D_k$，D_k 为第 k 个特征空间的维数。

表 1 以上主要算法的性能分析与比较

模型或算法	适用性	时间复杂度	优点	缺点	典型应用
k-means	凸分布	O(knl)	执行速度快；全局收敛；适合大数据	不能有效分类非凸数据集；对不适当的初始化，存在局部收敛情况	聚类分析，图像分割
势函数	任意形状	O(mn+km)	全局收敛	对大数据速度较慢	聚类相关应用
熵函数法 LBFC	凸分布	O(knl)	全局收敛	对大数据速度较慢容易产生重叠聚类	聚类相关应用
核 k-means	任意形状	O(knl)	执行速度快；全局收敛；非线性划分；适合大数据	同 k-means	非线性聚类分析相关应用
多核 多核聚类 多核多类 MMC	任意形状	近似 O(n) $O\left(\frac{D^{3.5}+nD}{\varepsilon^2}+\frac{D}{\varepsilon^4}\right)$	比单核 SVM 分类精度高；非线性划分；适合大数据	模型计算复杂	大数据聚类分析
GMM+EM	凸分布	O(knl)	执行速度较快；全局收敛	EM 速度受初始化影响较大；非凸数据集分割效果不好	聚类分析，图像分割，视频分析
谱聚类 NJW	任意形状	$O(n^3)$	关注数据之间的相似性，与对象属性无关；适合非凸数据集；实现了高维数据的降维	不适合大数据，大尺寸图像	聚类分析，小尺寸图像分割
Nyström 方法		$O(r^3+rn)$	适合大数据	模型计算复杂	图像分割，视频分析

结束语 概率统计为机器学习方法提供了一套有效的数学分析理论，可以用来建立和分析新型的机器学习模型和算法。计算机领域常用的统计分析模型和方法较为广泛，如最小二乘法、主成分分析、k-means 聚类、密度估计、回归的光滑

等。本文主要阐述了当今一些经典的以及较新的数据聚类模型，这些模型在本质上与统计分析基础具有一定联系，这些基础包括建立在概率测度上的 Hilbert 空间理论、Hilbert 空间中的积分算子学习特征函数问题、抽样技术、小样本条件下的近似理论、贝叶斯推断、信息熵、Gibbs 随机场及势函数、用于建模的各种统计模型。同时，文中介绍了所提的统计聚类模型的一些数学来源、原理、性质与推导，以及最新的应用。

当然，统计聚类模型远远不止这些，从广义上来讲，许多经典的统计模型可以用来分类数据，一些已经得到使用，如判别分析、提升和装袋(boosting and bagging)、投影寻踪、集成方法。它们在应用领域的研究更为广泛，在此不一一讨论。目前，一些有趣的研究方向可以深入研究，如空间约束条件下的 GMM 模型及其应用、核函数在经典统计模型中的扩展与应用、降低谱聚类复杂性的方法、多尺度谱聚类、Hilbert 空间积分算子的扩展、无限高斯混合模型、机器学习中的高斯过程、狄利克莱先验或过程、多项式先验分布等等。

参 考 文 献

[1] Theodoridis S, Koutrombas K. Pattern Recognition [M]. Beijing: China Machine Press, 2003

[2] Duda R O, Hart P E, Stork D G. Pattern Classifi -cation[M]. Beijing: China Machine Press, 2003

[3] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. NEW YORK: Plenum Press, 1981

[4] Selim S Z, Ismail M A. On the local optimality of the fuzzy isodata clustering algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986: 8(2): 284-288

[5] Hathaway R J, Bezdek J C. Local convergence of the fuzzy c-Means algorithms[J]. Pattern Recognition, 1986, 19(6): 477-480

[6] Chaturvedi A, Foods K, Green P E, et al. K-modes clustering [J]. Journal of Classification, 2001(18): 35-55

[7] Pal N R, Pal K, Bezdek J C. A mixed c-means clustering model [C]// Proceedings of the Sixth IEEE International Conference on Fuzzy Systems. Barcelona, Spain, 1997: 11-21

[8] Belacel N , Hansen P, Mladenovic N. Fuzzy J-Means: a new heuristic for fuzzy clustering [J]. Pattern Recognition, 2002, 35: 2193-2200

[9] Chepoi V, Dumitrescu D. Fuzzy clustering with structural constraints[J]. Fuzzy Sets and Systems, 1999, 105: 91-97

[10] Likas A, Vlassis N, Verbeekb J. The global k-means clustering algorithm[J]. Pattern Recognition, 2003, 36: 451-461

[11] Davé R N, Krishnapuram R. Robust clustering methods: a unified view[J]. IEEE Trans. on Fuzzy Systems, 1997, 5(2): 270-293

[12] Yager R R, Filev D P. Approximate clustering via the mountain method[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1994, 24(8): 1279-1284

[13] Yao J, Dash M, Tan S T, et al. Entropy-based fuzzy clustering and fuzzy modeling[J]. Fuzzy Sets and Systems, 2000, 113(3): 381-388

[14] Karayiannis N B. MECA: Maximum entropy clustering algorithm [C]// Proceedings of the Third IEEE Conference on Fuzzy Systems. Orlando, FL, 1994: 630-635

[15] Li Rui-ping, Mukaidono M. Gaussian clustering method based on maximum-fuzzy-entropy interpretation[J]. Fuzzy Sets and Systems, 1999, 102(2): 253-258

[16] Gerardo B, Liu Xiao-min. A least biased fuzzy clustering method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(9): 954-960

[17] Schneider A. Weighted possibilistic c-means clustering algorithms[C]// The Ninth IEEE International Conference on Fuzzy Systems. Vol. 1, San Antonio, 2000: 176-180

[18] 张志华，郑南宁，史罡. 极大熵聚类算法及其收敛性分析[J]. 中国科学(E辑)，2001，31(1)：59-70

[19] Davé R N, Krishnapuram R. Robust clustering methods: a unified view[J]. IEEE Transactions on Fuzzy Systems, 1997, 5(2): 270-293

[20] Gokcay E, Principe J C. Information theoretic clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 158-171

[21] Gunter S, Schraudolph N N, Vishwanathan S V N. Fast iterative kernel principal component analysis [J]. Journal of Machine Learning Research, 2007, 8: 1893-1918

[22] Asa B-H, et al. Support vector clustering[J]. Journal of Machine Learning Research, 2001, 2: 125-137

[23] Zhang Dao-qiang, Chen Song-can. A novel kernelized fuzzy C-means algorithm with application in medical image segmentation [J]. Artificial Intelligence in Medicine, 2004, 32: 37-50

[24] Girolami M. Mercer kernel-based clustering in feature space[J]. IEEE Transactions on Neural Networks, 2002, 13(3): 780-784

[25] Rakotomamonjy A, et al. Simple MKL [J]. Journal of Machine Learning Research, 2008, 9: 2491-2521

[26] Zhao Bin, Kwok J T, Zhang Chang-shui. Multiple kernel clustering[C]// Proceedings of the SIAM International Conference on Data Mining. 2009: 638-649

[27] Scholköpf B, Smpla A, Muller K-R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation, 1998, 10: 1299-1319

[28] Dhillon I S, Guan Yu-qiang, Kulis B. Weighted graph cuts without eigenvectors: a multilevel approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(11): 1-14

[29] Filippone M, Camastra F, Masulli F, et al. A survey of kernel and spectral methods for clustering [J]. Pattern Recognition, 2008, 41(1): 176-190

[30] Sonnenburg S. Large scale multiple kernel learning[J]. Journal of Machine Learning Research, 2006, 7: 1531-1565

[31] Kloft M, Brefeld U, Laskov P, et al. Non-sparse Multiple Kernel Learning[C]// NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels. Whistler, Canada, 2008: 1-4

[32] Subrahmanya N, Shin Y C. Sparse multiple kernel learning for signal processing applications[J]. IEEE Transactions on Pattern

Analysis and Machine Intelligence,2010,32(5):788-798

[33] Szafranski M,Grandvalet Y,Rakotomamonjy A. Composite kernel learning[J]. Machine Learning,2010,79:73-103

[34] Lanckriet G,et al. A statistical framework for genomic data fusion[J]. Bioinformatics,2004,20:2626-2635

[35] 高琰,谷士文,唐璀,等. 机器学习中谱聚类方法的研究[J]. 计算机科学,2007,34(2):201-203

[36] Ng A,Jordan M,Weiss Y. On spectral clustering:analysis and an algorithm[J]. Neural Information Processing Systems,2001,14:849- 856

[37] Weiss Y. Segmentation using eigenvectors:a unifying view[C]// International Conference on Computer Vision. 1999:1-7

[38] Meila M,Shi Jianbo. Learning segmentation by random walks [J]. Neural Information Processing Systems,2002,13:837-879

[39] Fook K T. Multilevel spectral clustering:graph partitions and image segmentation[D]. Cambridge:Massachusetts Institute of Technology,2008

[40] 张向荣,等. 基于免疫谱聚类的图像分割[J]. 软件学报,2010,21(9):2196-2205

[41] 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

[42] Zelnik-Manor L,Perona P. Self-tuning spectral clustering[R]. Neural Information Processing Systems,2004:1601-1608

[43] 李小斌,田铮. 基于谱聚类的图像多尺度随机树分割[J]. 中国科学 E 辑:信息科学,2007,37(8):1073- 1085

[44] Bengio Y,et al. Spectral clustering and kernel PCA are learning eigenfunctions[R]. TR1239. Montréal:Département d'Informatique et Recherche Opérationnelle,Université de Montréal,2003

[45] Fowlkes C,Belongie S,Chung F,et al. Spectral grouping using the Nyström method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(2):214-225

[46] Bilmes J A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models[R]. TR-97-021. Berkeley:Computer Science Division,Department of Electrical Engineering and Computer Science U. C. Berkeley,1998

[47] Bishop C M. Neural Networks for Pattern Recognition[M]. Oxford:Clarendon Press,1995

[48] Sanjay-Gopal S, Hebert T J. Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm[J]. IEEE Transactions on Image Processing,1998,7(7):1014-1028

[49] Blekas K,et al. A spatially constrained mixture model for image segmentation[J]. IEEE Transactions on Neural Networks,2005,16(2):494-498

[50] Nikou C, et al. A Bayesian framework for image segmentation with spatially varying mixtures[J]. IEEE Transactions on Image Processing,2010,19(9):2278-2289

[51] Nikou C,et al. A class-adaptive spatially variant mixture model for image segmentation[J]. IEEE Transactions on Image Processing,2007,16(4):1121-1130

[52] Nguyen T M,et al. An extension of the standard mixture model for image segmentation[J]. IEEE Transactions on Neural Networks,2010,21(8):1326-1338

[53] 马江洪,葛咏. 图像线状模式的有限混合模型及其 EM 算法[J]. 计算机学报,2007,30(2):288-296

[54] 向日华,王润生. 一种基于高斯混合模型的距离图像分割算法[J]. 软件学报,2003,14(7):1250-1257

[55] Myronenko A,Song Xu-bo. Point set registration:coherent point drift[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(12):2262-2275

(上接第 17 页)

[14] Hu W,Li J H,Shi J J. A Novel Approach to Cyberspace Security Situation Based on the Vulnerabilities Analysis[A]// Proceedings of the 6th World Congress on Intelligent Control and Automation[C]. vol 1,Dalian,China,2006:4747-4751

[15] Zhang Y,Tan X B,Xi H S. A Novel Approach to Network Security Situation Awareness Based on Multi-perspective Analysis [A]// IEEE 2007 International Conference on Computational Intelligence and Security[C]. Harbin,China,2007:768-772

[16] 陈秀真,郑庆华,管晓宏,等. 层次化网络安全威胁态势量化评估方法[J]. 软件学报,2006,17(4):885-897

[17] Chen X Z,Zheng Q H,Guan X H,et al. Quantitative hierarchical threat evaluation model for network security[J]. Journal of Software,2006,17(4):885-897

[18] 韦勇,连一峰,冯登国. 基于信息融合的网络安全态势评估模型[J]. 计算机研究与发展,2009,46(3):353-362

[19] Wei Y,Lian Y F,Feng D G. A Network Security Situational Awareness Model Based on Information Fusion[J]. Journal of Computer Research and Development,2009,46(3):353-362

[20] Han J W,Pei J,Yin Y W. Mining frequent patterns without candidate generation[J]. ACM SIGMOD Record,2000,29(2):1-12

[21] Mika K. A knowledge discovery methodology for telecommunication network alarm databases[D]. Finland,University of Helsinki,1999

[22] Dempster A. Upper and lower probabilities induced by multivalued mapping[J]. Annals of Mathematical Statistics,1967,38(2):325-339

[23] Kedar-Cabelli S T,McCarty L T. Explanation-based generalization as resolution theorem proving[A]// Proceedings of the Fourth International Workshop on Machine Learning[C]. San Mateo,CA,1987:383-389

[24] Mierswa I,Klinkberg R,Fischer S,et al. A flexible platform for knowledge discovery experiments:YALE-yet another learning environment[A]//LLWA 03 Tagung. GI-Worksh. Woche Lern. Leh. Wiss. Adapt[C]. Karlsruhe,Germany,2003

[25] Haines J W,Lippmann R P,Fried D J,et al. DARPA intrusion detection system evaluation:Design and procedures[R]. 1062. Lexington:MIT Lincoln Laboratory,1999

[26] DARPA Intrusion Detection Evaluation[EB/OL]. http://www.ll. mit. edu/mission/communications/ist/corpora/ideval/index.html