核聚类算法

引言

聚类是将一组给定的未知类标号的样本分成内在的多个类别，使得同一类中的样本具有较高的相似度，而不同类中的样本差别大。聚类分析的目的是揭示和刻画数据的内在结构，其内容涉及统计学、生物学、以及机器学习等研究领域，并在模式识别、数据分析和挖掘、图像处理等领域获得了广泛的应用[JD88][JMF99] [JDM00]。常用的聚类方法可分为如下几类 [HK00]：划分方法，层次聚类方法，基于密度的方法，基于网格的方法和基于模型的方法。本章侧重于第一种方法，即划分聚类方法的研究。

最著名与最常用的划分聚类方法是 C-均值 [Mac67] 及其推广模糊 C-均值（Fuzzy C-Means，简称为 FCM）[Bez81] 算法。C-均值算法首先由 J. MacQueen [Mac67] 提出，其原理是：首先定义一个准则函数，并随机选择 C 个初始聚类中心，然后根据样本与聚类中心的距离，将该样本划分到该类中；再重新计算每个类的聚类中心。此过程不断重复，直到准则函数最小。通常，准则函数选为样本和聚类中心的平方误差的总和。C-均值算法的缺点是：1）准则函数是不可微的，不能直接应用无约束最优化的梯度方法，导致算法训练没有一个终止准则，结果严重依赖初始聚类中心的选取；2）由于采用平方误差和准则，该方法仅适合于发现球形或类似球形分布的类别；3）难以发现大小差别很大的类别；4）对噪声和野值（Outlier）敏感；5）类别数 C 必须要事先确定。此后，J. C. Bezdek [Bez81] 提出了里程碑式的模糊 C-均值算法，通过引入样本到聚类中心的隶属度，使准则函数不仅可微，而且软化了模式的归属，由此解决了第一个问题。为了解决第二个问题，许多学者在 FCM 算法的基础上，通过修改准则函数，从而达到对不同形状分布样本的聚类。例如，适合椭球形分布样本聚类的 Gustafson-Kessel 算法 [GK79]，适合线形分布样本聚类的模糊 C-方差算法 [Bez81]，以及适合环形分布样本聚类的模糊 C-球壳算法 [Dav90]。除此之外，R. O. Duda 和 P. E. Hart [DH73] 还提出用类内散度矩阵的行列式作为准则函数，R. Krishnapuram 等人 [KK00] 给出了此准则函数下的迭代求解公式。K. L. Wu 和 M. S. Yang [WY02] 最近又提出一种指数型准则函数，得到的算法可以有效发现大小差别很大的类别，并且对噪声不敏感。

上述大多数算法是由修改 FCM 中的准则函数得到，其缺点是往往只能对某一种分布形式的样本有效聚类，并且和 FCM 算法一样，对噪声很敏感。Y. Ohashi [Oha84] 首先对 FCM 的噪声敏感性问题进行研究，提出修改的鲁棒型准则函数。

随后，R. N. Dave [Dav91]独立地提出噪声聚类算法（记为 NC），在一定条件下，NC 算法和 Ohashi 提出的算法是等价的，它们仅适合于只有一种噪声的情形[DK97]。R. Krishnapuram 和 J. M. Keller [KK93][BCM96][KK96] 提出著名的可能性 C-均值聚类算法（Possibilistic C-Means，简称为 PCM），其放宽了对隶属度函数的约束，对噪声有一定程度的鲁棒性。在样本只有一类时，PCM 和 NC 是等价的。PCM 可以处理多噪声情形，其缺点是对初始条件太敏感，一般用 FCM 先得到一个初始的估计，但当噪声比较严重时，此方法失效。

本章把核方法用于无监督聚类，提出了一系列核聚类算法：首先在 5.2 节讲述两种模糊核 C-均值算法（KFCM-Ⅰ和 KFCM-Ⅱ）；5.3 节介绍可能性核 C-均值算法（KPCM）；5.4 节讲述一种联机的核聚类算法（ROC）。在人工和 Benchmark 数据集上的结果显示，上面所提到的核聚类算法是鲁棒的，适合对不完整或缺失数据、包含噪声和野值数据的聚类。5.5 节介绍了核聚类算法在不完整数据集上聚类的应用，最后在 5.6 节对本章做了一个小节。

5.2 两种模糊核聚类算法

本章侧重于软聚类（模糊 C-均值——FCM），但其描述手段同样适合于硬聚类（HCM）等同类问题。

5.2.1 问题的刻画

FCM 是由 J. C. Bezdek [Bez81] 从硬 C-均值算法（记为 HCM）推广而来，已成为最常用和讨论较多的聚类算法之一。其描述如下：

令 $X=\{x_i, i=1,2,...,n\}$ 是一训练样本集， $X\subseteq R^p$， c 为预定的类别数目， v_i $(i=1,2,...,c)$ 为第 i 个聚类的中心，u_{ik} $(i=1,2,...,c, k=1,2,...,n)$ 是第 k 个样本对第 i 类的隶属度函数，且 $0\le u_{ik}\le 1$ 及 $0<\sum_{k=1}^{n}u_{ik}<n$，FCM 的目标函数为：

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^m\|x_k-v_i\|^2$$

(5.1)

其中，$U=\{u_{ik}\}$， $v=(v_1,v_2,...,v_c)$， $m>1$ 为常数，其约束为

$$\sum_{i=1}^{c}u_{ik}=1,\forall k=1,2,...,n$$

(5.2)

在约束（5.2）下优化（5.1）式得：

$$u_{ik} = \frac{(1/\| x_k - v_i \|^2)^{1/(m-1)}}{\sum_{j=1}^{c}(1/\| x_k - v_j \|^2)^{1/(m-1)}}, \forall i = 1,2,...,c, k = 1,2,...,n$$

(5.3)

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}, \forall i = 1,2,...,c$$

(5.4)

直接计算U与v存在困难，J. C. Bezdek [Bez81] 利用交替优化算法或形如 ***z=f(z)***方程的不动点算法有效地求解了U与v，即U、v的交替迭代求解收敛到（5.1）式的局部最小点。分析与实验已证实 FCM 是 HCM 的推广，且聚类性能优于 HCM，但两者存在的一个共同不足是仅适用于球状或椭球状聚类，且对噪声及其野值（Outlier）极为敏感。下文中可以看到，通过把核方法引入聚类可以有效解决上述问题。

5.2.2 特征空间中的模糊核聚类算法（KFCM-Ⅰ）

在特征空间中进行聚类包含两个步骤：首先通过一个非线性映射$\Phi : c \to F$（$x \in R^p \to \Phi(x) \in R^q$，$q > p$，甚至可以为无穷维）将输入空间$c$变换至高维特征空间$F$；然后在特征空间$F$中进行聚类。这里的一个关键观察是由（5.4）式，聚类中心可由样本集线性组合表示，这一表示即称为对偶形式表示。

记F中的聚类中心v_i的对偶表示为

$$v_i = \sum_{k=1}^{n} b_{ik}\Phi(x_k), \forall i = 1,2,...,c$$

(5.5)

则特征空间中的模糊核聚类算法（记为 KFCM-Ⅰ）的目标函数为：

$$J_m(U,v) = J_m(U, b_1, b_2,..., b_c) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \| \Phi(x_k) - \sum_{l=1}^{n} b_{il}\Phi(x_l) \|^2$$

(5.6)

其中$b_i=(b_{i1},b_{i2},...,b_{in})^T, i=1,2,...,c$，（5.6）式中

$$
\begin{aligned}
&\|\Phi(x_k)-\sum_{l=1}^{n} b_{il}\Phi(x_l)\|^2 \\
&=\Phi(x_k)^T\Phi(x_k)-2\sum_{l=1}^{n} b_{il}\Phi(x_k)^T\Phi(x_l)+\sum_{l=1}^{n}\sum_{j=1}^{n} b_{il}\Phi(x_k)^T b_{ij}\Phi(x_j)
\end{aligned}
$$

(5.7)

由上式中的计算均以F中元素的内积形式出现，由核代入技巧知，上述内积定义了F中的一个核函数$K(x,y)$，满足$K(x,y)=\Phi(x)^T\Phi(y)$。反之，若某一核函数$K(x,y)$满足 Mercer 条件 [Mer09]，则它可诱导出一个映射，实现从某一低维输入空间到高维特征空间的隐映射。将$K(x,y)$代入（5.7）及（5.6）得：

$$J_m(U,b_1,b_2,...,b_c)=\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m(K_{kk}-2b_i^TK_k+b_i^TKb_i)$$

(5.8)

其中 $K_{ij}=K(x_i,x_j), i,j=1,2,...,n$ ， $K_k=(K_{k1},K_{k2},...,K_{kn})^T, k=1,2,...,n$ ，$K=(K_1,K_2,...,K_n), k=1,2,...,n$。

（5.8）式在（5.2）式的约束下经优化可得：

$$u_{ik}=\frac{(1/(K_{kk}-2b_i^TK_k+b_i^TKb_i))^{1/(m-1)}}{\sum_{j=1}^{c}(1/(K_{kk}-2b_j^TK_k+b_j^TKb_j))^{1/(m-1)}}, \forall i=1,2,...,c, k=1,2,...,n$$

(5.9)

$$b_i=\frac{\sum_{k=1}^{n}u_{ik}^mK^{-1}K_k}{\sum_{k=1}^{n}u_{ik}^m}, \forall i=1,2,...,c$$

(5.10)

由此可得 KFCM-Ⅰ的交替迭代算法如下：

Step1：设定聚类数目c和参数m。

Step2：初始化各个系数向量b_i，计算核矩阵K及其逆矩阵K^{-1}。

Step3：重复下面的运算，直到各个样本的隶属度值稳定：

（a）：用当前的系数向量根据式（5.9）更新隶属度，

（b）：用当前的隶属度根据式（5.10）更新各个系数向量。

求得 b_i 后，由式（5.5）即可得出特征空间中聚类中心的表达式，此时由于 Φ 未知，所以不能得到聚类中心的具体值，但可由（5.7）式求出其和样本间的距离。事实上，知道了样本和聚类中心间的距离，就可以判别样本属于哪一类了。

此算法中利用聚类中心的对偶表示进而获得在特征空间中的聚类，类似于 SVM 中分类超平面的求解，是 FCM 的一种自然推广。其不足一是 F 中的聚类中心无法在输入空间中加以描述，原因之一是（5.5）式未必存在原像，从而失去了原 FCM 直观的优点。二是（10）式中矩阵求逆所导致的速度下降。鉴于此，下一小节给出了输入空间中的模糊核聚类算法（KFCM-Ⅱ）。

5.2.3 输入空间中的模糊核聚类算法（KFCM-Ⅱ）

定义输入空间中的模糊核聚类算法（记为 KFCM-Ⅱ）的目标函数为：

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m}\|\Phi(x_k)-\Phi(v_i)\|^2 \qquad (5.11)$$

其中 v_i 为输入空间中的聚类中心（$i=1,2,...,c$），类似（5.6）式的展开并进行核代入，有

$$\|\Phi(x_k)-\Phi(v_i)\|^2=K(x_k,x_k)+K(v_i,v_i)-2K(x_k,v_i) \qquad (5.12)$$

由此可以定义下式

$$d(x,y)\overset{\Delta}{=}\|\Phi(x)-\Phi(y)\| \qquad (5.13)$$

事实上，$d(x,y)$ 为特征空间中的欧氏距离，核代入使之在原输入空间中诱导出了一类核依赖的新的距离度量，这是核方法带来的新观点，由此将 FCM 在欧氏距离下的执行推广到了同一空间中不同距离度量的新的聚类。

将（5.12）式代入（5.11）式，在（5.2 式）的约束下优化（5.11）式可得：

$$u_{ik}=\frac{(1/(K(x_k,x_k)+K(v_i,v_i)-2K(x_k,v_i)))^{1/(m-1)}}{\sum_{j=1}^{c}(1/(K(x_k,x_k)+K(v_j,v_j)-2K(x_k,v_j)))^{1/(m-1)}} \qquad (5.14)$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}^{m}\tilde{K}(x_k,v_i)x_k}{\sum_{k=1}^{n}u_{ik}^{m}\tilde{K}(x_k,v_i)} \qquad (5.15)$$

其中，对高斯核函数、多项式核函数和 Sigmoid 核函数，（5.15）式中的 $\tilde{K}(x_k, v_i)$ 分别为：

（ⅰ）高斯核函数，有 $\tilde{K}(x_k, v_i) = K(x_k, v_i)$；

（ⅱ）多项式核函数，有 $\tilde{K}(x_k, v_i) = \left(\frac{K(x_k, v_i)}{K(v_i, v_i)} \right)^{\frac{d-1}{d}}$；

（ⅲ）Sigmoid 核函数，有 $\tilde{K}(x_k, v_i) = \frac{(1-K(x_k, v_i))}{(1-K(v_i, v_i))} \cdot \frac{(1-K(x_k, v_i))}{(1-K(v_i, v_i))}$。

以多项式核函数为例证明。（5.11）式对 v_i 求导，并令其为 0，得

$$\begin{aligned}\frac{\partial J_m}{\partial v_i} &= \sum_{k=1}^{n} u_{ik}^m \left(\frac{\partial K(v_i, v_i)}{\partial v_i} - 2\frac{\partial K(x_k, v_i)}{\partial v_i}\right) \\ &= \sum_{k=1}^{n} u_{ik}^m \left(d \cdot (v_i^T v_i + c)^{d-1} \cdot 2v_i - 2d \cdot (x_k^T v_i + c)^{d-1} \cdot x_k\right) = 0\end{aligned} \tag{5.16}$$

从（5.16）式可以解得 v_i，对比（5.15）式，可知 $\tilde{K}(x_k, v_i) = \left(\frac{K(x_k, v_i)}{K(v_i, v_i)} \right)^{\frac{d-1}{d}}$。

由（5.15）式可知，v_i 仍属于输入空间，但对各 x_k 的加权不同于 FCM 中的（5.4）式，其实质是 FCM 将几乎同样的隶属度赋给了远离每一聚类中心的点，即使它们对每一聚类具有不同程度的隶属度。由此造成了 FCM 在噪声环境下的失效。而 KFCM-Ⅱ尽管与 FCM 采用了相同的（5.2）式，但事实上由于（5.15）式加权系数中 $\tilde{K}(x_k, v_i)$ 的加入，使其对噪声点和野值赋予了不同的但是直觉上合理的权值，这一点在当核函数取高斯函数时解释更为直观。

KFCM-Ⅱ算法描述如下：

Step1：设定聚类数目 c 和参数 m。

Step2：初始化各个聚类中心 v_i。

Step3：重复下面的运算，直到各个样本的隶属度值稳定：

（a）：用当前的聚类中心根据式（5.14）更新隶属度，

（b）：用当前的聚类中心和隶属度根据式（5.15）更新各个聚类中心。

由于（5.14）式与（5.15）式中无矩阵求逆，致使 KFCM-Ⅱ算法的执行比 KFCM-Ⅰ快速。

5.2.4 KFCM-Ⅱ算法的鲁棒性分析

一个好的聚类算法应是鲁棒的，即能容忍噪声和野值，如此才具有实际的应用价值。本小节借助影响函数分析［Hub81］证明上小节提出的 KFCM-Ⅱ算法是 M-鲁棒估计的。一个度量是鲁棒的，在数学上意味着其对应的影响函数有界。为此，在本节中将证明上述核诱导的度量是鲁棒的。为表述方便，本节仅限于讨论单变量数据。

设$\{x_i, i = 1,2,\ldots,n\}$是样本数据集，$q$ 是一要估计的参数，定义如下函数

$$L(q) = \sum_{k=1}^{n} r(x_k - q) \tag{5.16}$$

其中 r 是一个任意函数，对 $L(q)$ 关于q 求导并令导数为 0，即得到 M-估计

$$\frac{\partial L(q)}{\partial q} = \sum_{k=1}^{n} \frac{\partial r(x_k - q)}{\partial q} = \sum_{k=1}^{n} j(x_k - q) \tag{5.17}$$

其中$j(x-q) = \dfrac{\partial}{\partial q} r(x-q)$。当 $r(x-q) = (x-q)^2$时，M-估计即为样本均值，它等价于传统的最小平方估计；若 $r(x-q) = |x-q|$，M-估计即为样本中值。

（5.17）式的解估计可表示为

$$\hat{q} = \sum_{k=1}^{n} \frac{w_k}{\sum_{k=1}^{n} w_k} w_k \tag{5.18}$$

其中$w_k = \dfrac{j(x_k - q)}{x_k - q}$，一般情形下（5.18）式难以直接求解，但可通过迭代法迭代求得。下面定义相应的影响函数$IF(x; F, q)$为

$$IF(x; F, q) = \frac{j(x-q)}{\int j'(x-q)\, dP_X(x)} \tag{5.19}$$

其中$P_X(x)$表示 X 的分布。影响函数 IF 用来评估单个数据对于估计的相对影响程度。若 IF 无界，则估计缺乏鲁棒性。现定义总误差敏感度$GES(r,q)$为

$$GES(r,q) = \sup_x |IF(x; r, q)| \tag{5.20}$$

$GES(r,q)$度量了对数据的微小扰动所产生的对估计的影响。若 GES 有界，

则估计是鲁棒的。下面证明相对于 KFCM-Ⅱ算法的目标函数（5.11）式，使用下面的核函数，所获得的参数$\{v_i\}$（(5.15）式)，估计是 M-鲁棒的。

（ⅰ）$K(x,y)=e^{-(x-y)^2/s^2}$

（ⅱ）$K(x,y)=e^{-|x-y|/s}$ 。

由 $r(x-q)=1-K(x,q)=1-K(x-q)$

对于（ⅰ），$j(x-q)=e^{-(x-q)^2/s^2}\cdot\frac{-2}{s^2}(x-q)$

对于（ⅱ），$j(x-q)=e^{-|x-q|/s}\cdot\frac{-1}{s}sign(x-q)$

对于（ⅰ）~（ⅱ）中的$j(x-q)$，有

$$\lim_{x\to\infty}j(x-q)=\lim_{x\to-\infty}j(x-q)=0 \tag{5.21}$$

即$j(x-q)$有界，其最大值存在且有限，因此$GES(r,q)$有限，说明 KFCM-Ⅱ的聚类中心的估计是鲁棒的。而对于欧氏距离，其 $r(x-q)=(x-q)^2$，$j(x-q)=2(q-x)$无界，$GES(r,q)=\infty$，所以 FCM 中聚类中心的估计不是鲁棒的。