第4章 模式识别—
4.3聚类分析

信控学院 蔡利梅

4.3.1概述

（1）基本概念

- 无监督模式识别

事先不知道要划分的是什么类别，没有类别已知的样本用来训练，通过某种方法直接把数据划分成若干类别，称为无监督模式识别、无监督学习、聚类分析等。

- 基于样本概率分布模型的聚类方法
- 基于样本间相似性度量的聚类方法

- 聚类中心

每类模式的聚集中心或具有代表性的模式，也称为标准模式。

- 聚类分析三要素

  - 模式相似性测度：衡量样本之间的相似性
  - 聚类准则函数：衡量样本集划分结果
  - 聚类算法：聚类的过程

（2）模式相似性测度

- 距离测度

  - 原理

    - 同类样本特征相似，不同类样本的特征显著不同时；同类样本会聚集在一个区域，不同类样本相对远离。
    - 样本点在特征空间距离的远近直接反映了相应样本所属类别，可作为样本相似性度量。
    - 距离越近，相似性越大，属于同一类的可能性就越大；距离越远，相似性越小，属于同一类的可能性就越小。

距离的定义

样本：$X_i=(x_{i1},x_{i2},\cdots,x_{in})^T$、$X_j=\left(x_{j1},x_{j2},\cdots,x_{jn}\right)^T$、$X_k=(x_{k1},x_{k2},\cdots,x_{kn})^T$，对任意两样本的距离定义为函数$d$，应满足：

1) $d\left(X_i,X_j\right)\geq 0$，当且仅当$X_i=X_j$时，等号成立；

2) $d\left(X_i,X_j\right)=d\left(X_j,X_i\right)$；

3) $d\left(X_i,X_j\right)\leq d\left(X_j,X_k\right)+d(X_i,X_k)$。

需要指出，模式识别中定义的某些距离测度不满足第3个条件，只是在广义意义上称之为距离。

常用距离

欧氏距离

$$d_e(X_i, X_j) = \| X_i - X_j \| = \sqrt{\sum_{k=1}^{n} | x_{ik} - x_{jk} |^2}$$

城市距离 Manhattan

$$d(X_i, X_j) = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right| \qquad k = 1, 2, \ldots, n$$

切氏距离 Chebyshev

$$d(X_i, X_j) = \max_{k} \left| x_{ik} - x_{jk} \right| \qquad k = 1, 2, \ldots, n$$

马氏距离 Mahalanobis

$$d^2 = (X - \mu)^T \Sigma^{-1} (X - \mu)$$

μ:均值向量; Σ: 协方差矩阵

明氏(Minkowski)距离

$$d\left(X_i, X_j\right)=\left[\sum_{k=1}^{n}\left|\left(x_{ik}-x_{jk}\right)\right|^{m}\right]^{1/m}$$ X_i, X_j均为n维模式向量

$$m=1: d_1\left(X_i, X_j\right)=\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|$$ 城市距离

$$m=2: d_2\left(X_i, X_j\right)=\left[\sum_{k=1}^{n}\left|\left(x_{ik}-x_{jk}\right)\right|^{2}\right]^{1/2}$$ 欧氏距离

$m \to \infty$: 切比雪夫距离

Camberra距离(Lance距离、Willims距离)

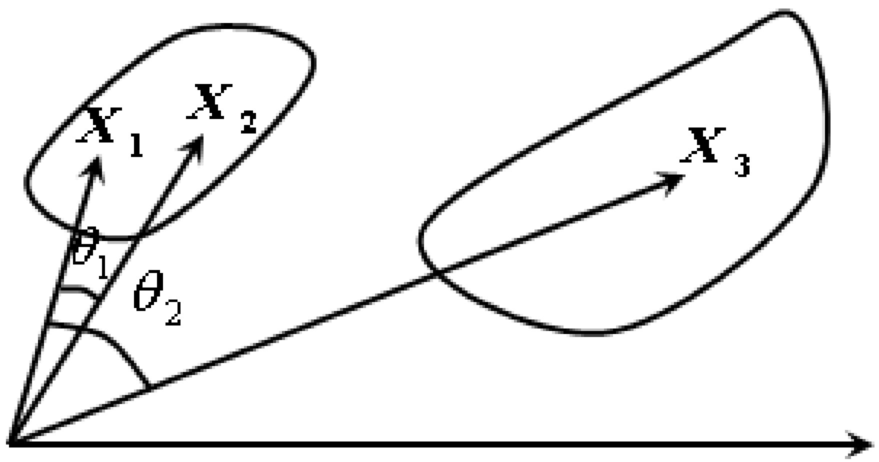
$$d\left(X_i, X_j\right)=\sum_{k=1}^{n}\frac{\left|x_{ik}-x_{jk}\right|}{\left|x_{ik}+x_{jk}\right|}, \quad \left(x_{ik}, x_{jk} \geq 0, x_{ik}+x_{jk} \neq 0\right)$$

相似测度

向量夹角余弦

$$s(X_i, X_j) = \frac{X_i^T X_j}{\|X_i\| \; \|X_j\|}$$

反映了几何相似性，模式向量具有扇形分布时常用



$$S(X_1, X_2) = \cos\theta_1$$
$$S(X_1, X_3) = \cos\theta_2$$

此时，余弦值越大，相似性越大

相关系数

$$r(X,Y)=\frac{(X-\mu_X)^T(Y-\mu_Y)}{[(X-\mu_X)^T(X-\mu_X)(Y-\mu_Y)^T(Y-\mu_Y)]^{1/2}}$$

指数相似系数

$$e(X_i,X_j)=\frac{1}{n}\sum_{k=1}^{n}\exp[-\frac{3(x_{ik}-x_{jk})^2}{4\sigma_k}]$$

其它相似测度

$$S(X_i,X_j)=\frac{\sum_k \min(x_{ik},x_{jk})}{\frac{1}{2}\sum_k (x_{ik}+x_{jk})}$$

$$S(X_i,X_j)=\frac{\sum_k \min(x_{ik},x_{jk})}{\sum_k \max(x_{ik},x_{jk})}$$

$$S(X_i,X_j)=\frac{\sum_k \min(x_{ik},x_{jk})}{\sum_k \sqrt{x_{ik}x_{jk}}}$$

4.3.2动态聚类

- 动态聚类方法的关键点
  - 采用距离度量样本间的相似性；
  - 确定某个评价聚类结果质量的准则函数；
  - 给定某个初始分类，通过迭代算法找出使准则函数取极值的最好聚类结果。

（1）K均值算法

- 基于最小误差平方和准则
- 原理：首先确定K个初始聚类中心，然后根据各类样本到聚类中心的距离平方和最小的准则，不断调整聚类中心，直到聚类合理。
- K需事先给定，不一定适合某些非监督学习问题

- 步骤

1）任选K个初始聚类中心$z_1(1), z_2(1), \cdots, z_K(1)$

2）逐个将每一样本按最小距离原则分配给K个聚类中心

$$\text{若} \|x - z_j(m)\| < \|x - z_i(m)\|, i = 1,2,\ldots, K, i \neq j, \ \text{则} x \in \omega_j(m)$$

3）计算新的聚类中心

$$z_i(m+1) = \frac{1}{N_i} \sum_{x \in \omega_i(m)} x \quad i = 1,2,\ldots, K$$

4）若$z_i(m+1) = z_i(m), i = 1, 2, \cdots, K$，算法收敛，否则，转到第2步，进行下一次迭代。

例：有20个二维样本，用K均值算法聚类。

$$X_1 \sim X_{20}: \left\{ \begin{matrix} (0\ \ 0)^T, (1\ \ 0)^T, (0\ \ 1)^T, (1\ \ 1)^T, (2\ \ 1)^T \\ (1\ \ 2)^T, (2\ \ 2)^T, (3\ \ 2)^T, (6\ \ 6)^T, (7\ \ 6)^T \\ (8\ \ 6)^T, (6\ \ 7)^T, (7\ \ 7)^T, (8\ \ 7)^T, (9\ \ 7)^T \\ (7\ \ 8)^T, (8\ \ 8)^T, (9\ \ 8)^T, (8\ \ 9)^T, (9\ \ 9)^T \end{matrix} \right\}$$

1）取K=2，令 $z_1(1) = X_1 = (0\ \ 0)^T \quad z_2(1) = X_2 = (1\ \ 0)^T$

2）样本归类

$$\|X_1 - z_1(1)\| < \|X_1 - z_2(1)\| \quad \therefore X_1 \in \omega_1(1)$$

$$\|X_2 - z_1(1)\| > \|X_2 - z_2(1)\| \quad \therefore X_2 \in \omega_2(1)$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

分配结果：

$$\omega_1(1) = \{X_1, X_3\}$$

$$\omega_2(1) = \{X_2, X_4, X_5, \cdots X_{20}\}$$

3）计算新的聚类中心

$$z_1(2) = \frac{1}{2}(X_1 + X_3) = (0 \quad 0.5)^T$$

$$z_2(2) = \frac{1}{18}(X_2 + X_4 + X_5 + \cdots + X_{20}) = (5.67 \quad 5.33)^T$$

4）判断算法是否收敛

$$\because z_1(2)\neq z_1(1)\quad z_2(2)\neq z_2(1)\quad \therefore \text{返回第二步}$$

2）重新分配样本

$$\|X_1-z_1(2)\|<\|X_1-z_2(2)\|\quad \therefore X_1\in\omega_1(2)$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

分配结果：

$$\omega_1(2)=\{X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8\}$$
$$\omega_2(2)=\{X_9,X_{10},X_{11},X_{12},\cdots,X_{20}\}$$

3）计算新的聚类中心

$$z_1(3)=\frac{1}{8}(X_1+\cdots+X_8)=(1.25\quad 1.13)^T$$

$$z_2(3)=\frac{1}{12}(X_9+X_{10}+\cdots+X_{20})=(7.67\quad 7.33)^T$$

4）判断算法是否收敛

$$\because z_1(2) \neq z_1(3) \quad z_2(2) \neq z_2(3) \quad \therefore \text{返回第二步}$$

2）重新分配样本，与上一次一致

$$\omega_1(2)=\{X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8\}$$
$$\omega_2(2)=\{X_9,X_{10},X_{11},X_{12},\cdots,X_{20}\}$$

3）计算新的聚类中心，与上一次一致

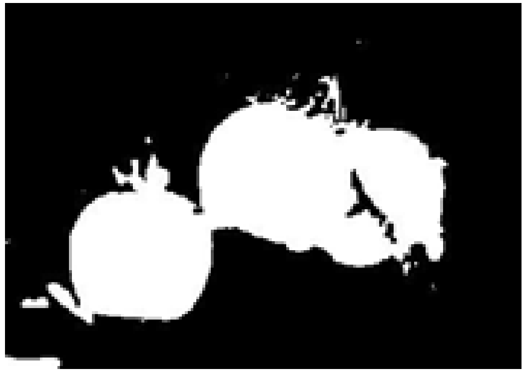
4）算法收敛

- 实例

对一幅苹果图像，利用色彩信息，实现聚类分割

原始图像

色调值作为聚类的数据；
初始聚类中心选择
0、1/6、2/6、3/6、4/6、5/6；



K均值聚类分割

（2）ISODATA算法

- Iterative Self-Organizing Data Analysis Techniques A：迭代自组织数据分析技术
- 原理
  - 与K均值算法相似，以均值迭代确定聚类中心
  - 可以调整参数，引入分裂与合并机制
    - 某两类中心间距小于某一阈值时，合并两类
    - 在某类样本标准差大于某一阈值时，或样本数目超过某一阈值时，分裂为两类
    - 类别数目少于某一阈值时，也实行分裂
  - 在类的样本数目少于某阈值时，可消除类

4.3.3高斯混合聚类

（1）基本概念

- 多元高斯分布

$$p(X)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right\}$$

- $X=[x_1\ \ x_2\ \ \cdots\ \ x_n]^T$，$n$维列向量
- $\mu=[\mu_1\ \ \mu_2\ \ \cdots\ \ \mu_n]^T=E\{X\}$：$n$维均值向量
- $\Sigma=E\{(X-\mu)(X-\mu)^T\}$：$n\times n$维协方差矩阵，$|\Sigma|$是$\Sigma$的行列式，$\Sigma^{-1}$是$\Sigma$的逆矩阵

- 高斯混合分布 $p_M(X)=\sum_{i=1}^{K}\alpha_i p(X\mid\mu_i,\Sigma_i)$

  - K个混合成分组成该分布，每个混合成分对应一个高斯分布，μ_i和Σ_i是第i个混合成分的参数
  - $\alpha_i \geq 0$，混合系数，$\sum_{i=1}^{K}\alpha_i = 1$，为选择第$i$个高斯混合成分的概率
  - 根据α_i定义的先验分布选择高斯混合成分，根据被选择的混合成分的概率密度函数进行采样，生成样本，进而构成样本集$A=\{X_1, X_2, \cdots, X_N\}$

- 令$z_j \in \{1,2,\cdots,K\}$表示生成样本X_j的高斯混合成分，z_j的先验概率$P(z_j=i)$对应于α_i，z_j的后验概率为

$$p_M(z_j=i \mid X_j)=\frac{P(z_j=i)\,p_M(X_j \mid z_j=i)}{p_M(X_j)}=\frac{\alpha_i p(X_j \mid \mu_i,\Sigma_i)}{\sum_{k=1}^{K}\alpha_k p(X_j \mid \mu_k,\Sigma_k)}$$

$$=\gamma_{ji},i=1,2,\cdots,K$$

- **高斯混合聚类**

假设数据服从高斯混合分布，K个高斯分布对应K个聚类簇，根据数据计算参数α_i、μ_i和Σ_i，再通过某种策略将样本归入某一类。

（2）参数求解

- 最大似然估计

样本集$A=\{X_1,X_2,\cdots,X_N\}$，样本服从高斯混合分布，寻找令似然函数$l(\theta)$或对数似然函数$H(\theta)=\ln l(\theta)$最大的参数$\hat{\theta}$。

$$l(\theta)=\prod_{j=1}^{N}\left(\sum_{i=1}^{K}\alpha_i p\left(X_j \mid \mu_i,\Sigma_i\right)\right)$$

可分为有监督和无监督两种情况

$$H(\theta)=\ln l(\theta)=\sum_{j=1}^{N}\ln\left(\sum_{i=1}^{K}\alpha_i p\left(X_j \mid \mu_i,\Sigma_i\right)\right)$$

$$=\sum_{j=1}^{N}\ln\left(\sum_{i=1}^{K}\alpha_i\frac{1}{(2\pi)^{n/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(X_j-\mu_i\right)^T\Sigma_i^{-1}\left(X_j-\mu_i\right)\right\}\right)$$

求最优

$$\because \theta = [\alpha_i, \mu_i, \Sigma_i]^T$$

$$\frac{\partial H}{\partial \mu_i} = \sum_{j=1}^{N} \frac{\alpha_i p(X_j \mid \mu_i, \Sigma_i)}{\sum_{i=1}^{K} \alpha_i p(X_j \mid \mu_i, \Sigma_i)} (X_j - \mu_i) = \sum_{j=1}^{N} \gamma_{ji} (X_j - \mu_i) = 0$$

$$\mu_i = \frac{\sum_{j=1}^{N} \gamma_{ji} X_j}{\sum_{j=1}^{N} \gamma_{ji}}$$

$$\Sigma_i = \frac{\sum_{j=1}^{N} \gamma_{ji} (X_j - \mu_i)(X_j - \mu_i)^T}{\sum_{j=1}^{N} \gamma_{ji}}$$

$$\alpha_i = \frac{\sum_{j=1}^{N} \gamma_{ji}}{N}$$

有约束条件，采用拉格朗日函数法

EM算法

期望最大化：Expectation Maximization，一种求参数的最大似然或最大后验概率估计的迭代方法。

- 定义分量数目K，对每个分量i设置α_i、μ_i和Σ_i的初始值，计算对数似然函数
- E步：计算样本X_j的对应z_j的后验概率γ_{ji}
- M步：利用前面推导的结论求α_i、μ_i和Σ_i
- 计算对数似然函数
- 判断参数是否收敛或对数似然函数是否收敛（如对数似然函数不再增长），收敛终止算法，不收敛返回E步

（3）样本归类

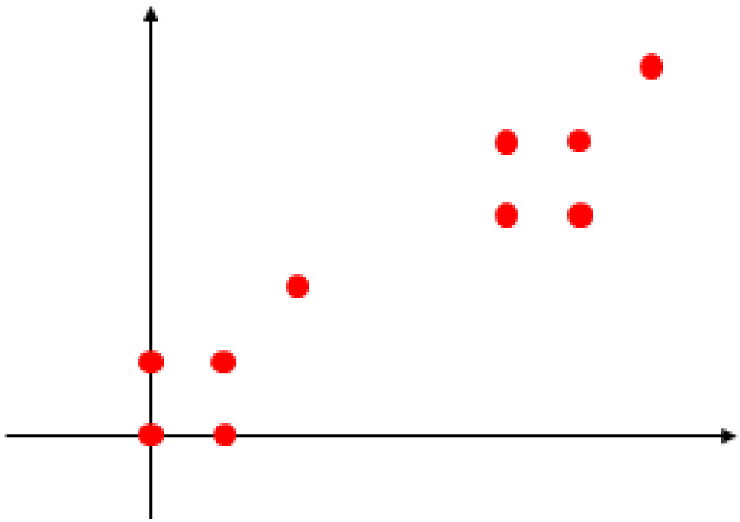
- 将样本X_j归入后验概率γ_{ji}最大的一类；
- 以α_i为概率随机选择K个高斯混合成分中的一个，将样本X_j代入高斯分布，判断输出概率是否大于阈值，不大于则重新选择。

例：有10个二维样本，进行高斯混合聚类。

$X_1=(0\quad 0)^T, X_2=(1\quad 0)^T, X_3=(2\quad 2)^T, X_4=(1\quad 1)^T, X_5=(0\quad 1)^T,$

$X_6=(5\quad 3)^T, X_7=(5\quad 4)^T, X_8=(6\quad 3)^T, X_9=(6\quad 4)^T, X_{10}=(7\quad 5)^T$

- **初始化**
  - 令K=2
  - $\mu_1=X_2, \mu_2=X_7$
  - $\alpha_1=\alpha_2=0.5,$ $\Sigma_1=\Sigma_2=I$



- 计算对数似然函数：$H(\theta) = \sum_{j=1}^{N} ln\left(\sum_{i=1}^{K} \alpha_i p\left(X_j \middle| \mu_i, \Sigma_i\right)\right)$ =-34.3078
- 计算样本由各混合成分生成的后验概率γ_{ji}

$$\gamma_{11} \approx 1, \gamma_{12} \approx 0, \gamma_{21} \approx 1, \gamma_{22} \approx 0,$$
$$\gamma_{31} \approx 1, \gamma_{32} \approx 0, \gamma_{41} \approx 1, \gamma_{42} \approx 0,$$
$$\gamma_{51} \approx 1, \gamma_{52} \approx 0, \gamma_{61} \approx 0, \gamma_{62} \approx 1,$$
$$\gamma_{71} \approx 0, \gamma_{72} \approx 1, \gamma_{81} \approx 0, \gamma_{82} \approx 1,$$
$$\gamma_{91} \approx 0, \gamma_{92} \approx 1, \gamma_{10,1} \approx 0, \gamma_{10,2} \approx 1。$$

- 计算新的参数

$$\mu_1 = (0.7994 \;\; 0.7994),\ \mu_2 = (5.7981 \; 3.7991)$$
$$\Sigma_1 = \begin{pmatrix} 0.5998 & 0.1991 \\ 0.1991 & 1.1986 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.6076 & 0.2019 \\ 0.2019 & 1.1999 \end{pmatrix}$$
$$\alpha_1 = 0.4998, \alpha_2 = 0.5002$$

- 计算对数似然函数： $H_{new}(\theta) = \sum_{j=1}^{N} ln\left(\sum_{i=1}^{K} \alpha_i p\left(X_j \middle| \mu_i, \Sigma_i\right)\right)$ =-29.7307
- 验证算法是否收敛$H_{new}(\theta) - H_{old}(\theta) > 1$，继续迭代
- 计算样本由各混合成分生成的后验概率γ_{ji}

$$\gamma_{11} \approx 1, \gamma_{12} \approx 0, \gamma_{21} \approx 1, \gamma_{22} \approx 0,$$
$$\gamma_{31} \approx 1, \gamma_{32} \approx 0, \gamma_{41} \approx 1, \gamma_{42} \approx 0,$$
$$\gamma_{51} \approx 1, \gamma_{52} \approx 0, \gamma_{61} \approx 0, \gamma_{62} \approx 1,$$
$$\gamma_{71} \approx 0, \gamma_{72} \approx 1, \gamma_{81} \approx 0, \gamma_{82} \approx 1,$$
$$\gamma_{91} \approx 0, \gamma_{92} \approx 1, \gamma_{10,1} \approx 0, \gamma_{10,2} \approx 1。$$

- 计算新的参数

$$\mu_1 = (0.8\ \ 0.8),\ \mu_2 = (5.8\ 3.8),\ \ \alpha_1 = 0.5,\ \alpha_2 = 0.5$$
$$\Sigma_1 = \begin{pmatrix} 0.56 & 0.36 \\ 0.36 & 0.56 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.56 & 0.36 \\ 0.36 & 0.56 \end{pmatrix}$$

- 计算对数似然函数： $H_{new}(\theta)=\sum_{j=1}^{N} ln\left(\sum_{i=1}^{K}\alpha_i p\left(X_j|\mu_i,\Sigma_i\right)\right)$=-26.8461
- 验证算法是否收敛$H_{new}(\theta)-H_{old}(\theta)>1$，继续迭代
- 计算样本由各混合成分生成的后验概率γ_{ji}

$$\gamma_{11}\approx 1,\gamma_{12}\approx 0,\gamma_{21}\approx 1,\gamma_{22}\approx 0,$$
$$\gamma_{31}\approx 1,\gamma_{32}\approx 0,\gamma_{41}\approx 1,\gamma_{42}\approx 0,$$
$$\gamma_{51}\approx 1,\gamma_{52}\approx 0,\gamma_{61}\approx 0,\gamma_{62}\approx 1,$$
$$\gamma_{71}\approx 0,\gamma_{72}\approx 1,\gamma_{81}\approx 0,\gamma_{82}\approx 1,$$
$$\gamma_{91}\approx 0,\gamma_{92}\approx 1,\gamma_{10,1}\approx 0,\gamma_{10,2}\approx 1。$$

- 计算新的参数

$\mu_1=(0.8\ \ 0.8)$, $\mu_2=(5.8\ 3.8)$， $\alpha_1=0.5,\alpha_2=0.5$

$$\Sigma_1=\begin{pmatrix}0.56 & 0.36\\ 0.36 & 0.56\end{pmatrix},\Sigma_2=\begin{pmatrix}0.56 & 0.36\\ 0.36 & 0.56\end{pmatrix}$$

- 计算对数似然函数：$H_{new}(\theta) = \sum_{j=1}^{N} ln\left(\sum_{i=1}^{K} \alpha_i p\left(X_j \middle| \mu_i, \Sigma_i\right)\right)$ =-26.8461
- 验证算法是否收敛$H_{new}(\theta) - H_{old}(\theta) < 1$，终止迭代
- 根据样本由各混合成分生成的后验概率γ_{ji}归类

ω_1：$\{X_1, X_2, X_3, X_4, X_5\}$，$\omega_2$：$\{X_6, X_7, X_8, X_9, X_{10}\}$