第5章 统计模式识别中的聚类方法

5.1 聚类分析

略

5.2 聚类准则

- 误差平方和准则函数 $J_e = \sum_{j=1}^{c} \sum_{k=1}^{n_j} ||X_k^j - m_j||^2$ ，其中 m_j 为类别 ω_j 的样本均值。
- 权平均平方距离和准则函数 $J_l = \sum_{j=1}^{c} P_j D_j^*$ ，其中 P_j 为各类的先验概率， $D_j^* = \frac{1}{c_{n_j}^2} \sum_{1 \le k,l \le n_j} ||X_K^j - X_l^j||^2$ ，其中C为组合数
- 类间距离和准则函数 J_b ，定义m为所有样本均值， $J_{b1} = \sum_{j=1}^{c} (m_j - m)^T (m_j - m)$ 和 $J_{b2} = \sum_{j=1}^{c} P_j (m_j - m)^T (m_j - m)$
- 离散度准则函数 $S_t = S_w + S_b$ ，在给定样本后是常数。
- 基于迹的准则函数：类内离散度 S_j 正比于沿各坐标轴方向的分量方差之和。 $J_{tw} = \mathrm{tr}[S_w]$， $J_{tb} = \mathrm{tr}[S_b]$
- 基于行列式的准则函数：其值正比于样本在各主轴上的相应方差之积。 $J_{dw} = |S_w|$， $J_{db} = |S_b|$ ，实际上尽量减少使用 $|S_b|$ 作为准则函数。
- 基于特征值的准则函数：应当使得 $S_w^{-1} S_b$ 的d个特征值的取值尽可能大。

5.3 基于分裂的聚类算法

1. 简单增类聚类算法：对每个类如果找不到聚类就自己当类中心。
2. 最大最小聚类算法，距离够远就算一个类中心，然后按最近邻分进去

5.4 基于合并的聚类算法

每一个都视为一类，然后逐渐减少，直到各类之间距离都大于一个值为止。

5.5 动态聚类算法

1. C-均值动态聚类算法（1）
   任选C个点作为聚类中心，对每个点按最近邻法分类，对每一类计算中心，再按最近邻法，依次进行，直到中心点不变。
2. C-均值动态聚类算法（2）
   同样任选，但是计算能得到 $\frac{n_i}{n_i - 1} ||X_{i,l}^k - Z_i^k||^2 - \frac{n_j}{n_j + 1} ||X_{i,l}^k - Z_j^k||^2$ 最大的i，j和k；若其小于0则退出，否则讲对应样本从i类挪到j类。其目的是使得 J_e 尽量小。
3. ISODATA算法
   类别数量不固定，参见课本P204
4. 基于样本和核的相似性度量的动态聚类算法
   在前两种中使用相似性计算。核函数可采用正态、主轴等。

5.6 基于近邻函数值准则的聚类算法

见课本P217

5.7 最小张树聚类算法

切掉最小生成树的最大权值边，直到所有的最大权值都小于一个阈值。

改进措施：计算主干上各点深度，若小于某个阈值则切断。