

# 第四章 特征选择和提取

# 第四章 特征选择和提取

- 特征选择和提取是模式识别中的一个关键问题
  - 前面讨论分类器设计的时候，一直假定已给出了特征向量维数确定的样本集，其中各样本的每一维都是该样本的一个特征；
  - 这些特征的选择是很重要的，它强烈地影响到分类器的设计及其性能；
  - 假若对不同的类别，这些特征的差别很大，则比较容易设计出具有较好性能的分类器。

# 第四章 特征选择和提取

- 特征选择和提取是构造模式识别系统时的一个重要课题
  - 在很多实际问题中，往往不容易找到那些最重要的特征，或受客观条件的限制，不能对它们进行有效的测量；
  - 因此在测量时，由于人们心理上的作用，只要条件许可总希望把特征取得多一些；
  - 另外，由于客观上的需要，为了突出某些有用信息，抑制无用信息，有意加上一些比值、指数或对数等组合计算特征；
  - 如果将数目很多的测量值不做分析，全部直接用作分类特征，不但耗时，而且会影响到分类的效果，产生“特征维数灾难”问题。

# 第四章 特征选择和提取

- 为了设计出效果好的分类器，通常需要对原始测量值集合进行分析，经过选择或变换处理，组成有效的识别特征；
- 在保证一定分类精度的前提下，减少特征维数，即进行“降维”处理，使分类器实现快速、准确和高效的分类。
- 为达到上述目的，关键是所提供的识别特征应具有很好的可分性，使分类器容易判别。为此，需对特征进行选择。
  - 应去掉模棱两可、不易判别的特征；
  - 所提供的特征不要重复，即去掉那些相关性很强且没有增加更多分类信息的特征。

# 第四章 特征选择和提取

- 说明

- 实际上，特征选择和提取这一任务应在设计分类器之前进行；
- 从通常的模式识别教学经验看，在讨论分类器设计之后讲述特征选择和提取，更有利于加深对该问题的理解。

# 第四章 特征选择和提取

- 所谓特征选择，就是从 $n$ 个度量值集合 $\{x_1, x_2, \dots, x_n\}$ 中，按某一准则选取出供分类用的子集，作为降维（ $m$ 维， $m < n$ ）的分类特征；
- 所谓特征提取，就是使 $(x_1, x_2, \dots, x_n)$ 通过某种变换，产生 $m$ 个特征 $(y_1, y_2, \dots, y_m)$  ( $m < n$ )，作为新的分类特征（或称为二次特征）；
- 其目的都是为了在尽可能保留识别信息的前提下，降低特征空间的维数，已达到有效的分类。

# 第四章 特征选择和提取

- 以细胞自动识别为例
  - 通过图像输入得到一批包括正常细胞和异常细胞的图像，我们的任务是根据这些图像区分哪些细胞是正常的，哪些细胞是异常的；
  - 首先找出一组能代表细胞性质的特征，为此可计算
    - 细胞总面积
    - 总光密度
    - 胞核面积
    - 核浆比
    - 细胞形状
    - 核内纹理
    - .....



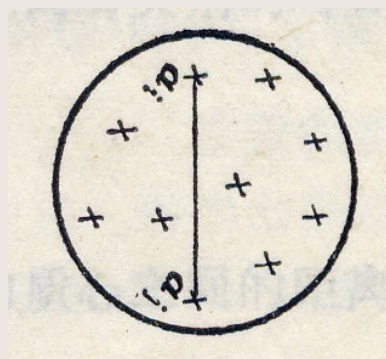
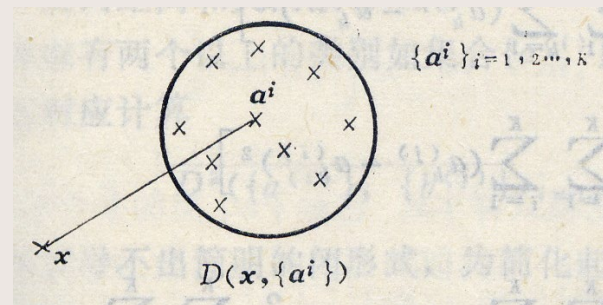
# 第四章 特征选择和提取

- 以细胞自动识别为例
  - 这样产生出来的原始特征可能很多（几十甚至几百个），或者说原始特征空间维数很高，需要降低（或称压缩）维数以便分类；
  - 一种方式是从原始特征中挑选出一些最有代表性的特征，称之为特征选择；
  - 另一种方式是用映射（或称变换）的方法把原始特征变换为较少的特征，称之为特征提取。



# 4.1 模式类别可分性的测度

- 距离和散布矩阵
  - 点到点之间的距离
  - 点到点集之间的距离
  - 类内距离



# 4.1 模式类别可分性的测度

---

- 距离和散布矩阵
  - 类内散布矩阵
  - 类间距离和类间散布矩阵
  - 多类模式集散布矩阵

## 4.2 特征选择

- 设有 $n$ 个可用作分类的测量值，为了在不降低（或尽量不降低）分类精度的前提下，减小特征空间的维数以减少计算量，需从中直接选出 $m$ 个作为分类的特征。
- 问题：在 $n$ 个测量值中选出哪一些作为分类特征，使其具有最小的分类错误？

## 4.2 特征选择

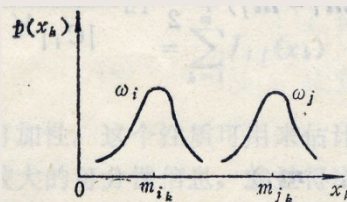
- 从n个测量值中选出m个特征，一共有  $C_n^m$  中可能的选法。
  - 一种“穷举”办法：对每种选法都用训练样本试分类一下，测出其正确分类率，然后做出性能最好的选择，此时需要试探的特征子集的种类达  $C_n^m = \frac{n!}{m!(n-m)!}$  种，非常耗时。
  - 需寻找一种简便的可分性准则，间接判断每一种子集的优劣。
    - 对于独立特征的选择准则
    - 一般特征的散布矩阵准则

## 4.2 特征选择

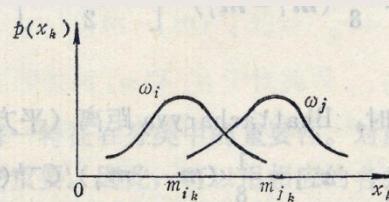
- 对于独立特征的选择准则
  - 类别可分性准则应具有这样的特点，即不同类别模式特征的均值向量之间的距离应最大，而属于同一类的模式特征，其方差之和应最小。
  - 假设各原始特征测量值是统计独立的，此时，只需对训练样本的 $n$ 个测量值独立地进行分析，从中选出 $m$ 个最好的作为分类特征即可。
    - 例：对于 $\omega_i$ 和 $\omega_j$ 两类训练样本的特征选择

## 4.2 特征选择

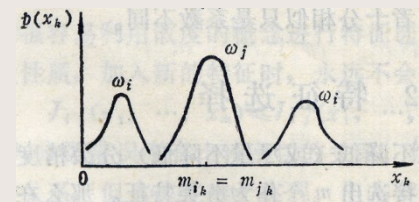
- 讨论：上述基于距离测度的可分性准则，其适用范围与模式特征的分布有关。
  - 三种不同模式分布的情况
    - (a) 中特征 $x_k$ 的分布有很好的可分性，通过它足以分离 $\omega_i$ 和 $\omega_j$ 两种类别；
    - (b) 中的特征分布有很大的重叠，单靠 $x_k$ 达不到较好的分类，需要增加其它特征；
    - (c) 中的 $\omega_i$ 类特征 $x_k$ 的分布有两个最大值，虽然它与 $\omega_j$ 的分布没有重叠，但计算 $G_k$ 约等于0，此时再利用 $G_k$ 作为可分性准则已不合适。
  - 因此，假若类概率密度函数不是或不近似正态分布，均值和方差就不足以用来估计类别的可分性，此时该准则函数不完全适用。



(a)



(b)



(c)

## 4.2 特征选择

- 一般特征的散布矩阵准则

- 类内、类间的散布矩阵 $S_w$ 和 $S_b$
- 类间离散度越大且类内离散度越小，可分性越好。
- 散布矩阵准则 $J_1$ 和 $J_2$ 形式
- 使 $J_1$ 或 $J_2$ 最大的子集可作为所选择的分类特征。
  - 注：这里计算的散布矩阵不受模式分布形式的限制，但需要有足够数量的模式样本才能获得有效的结果



# 作业

- 设有如下三类模式样本集 $\omega_1$ ,  $\omega_2$ 和 $\omega_3$ , 其先验概率相等, 求 $S_w$ 和 $S_b$

$$\omega_1: \{(1 \ 0)^T, (2 \ 0)^T, (1 \ 1)^T\}$$

$$\omega_2: \{(-1 \ 0)^T, (0 \ 1)^T, (-1 \ 1)^T\}$$

$$\omega_3: \{(-1 \ -1)^T, (0 \ -1)^T, (0 \ -2)^T\}$$

## 4.3 离散K-L变换

- 全称：Karhunen-Loeve变换（卡洛南-洛伊变换）
- 前面讨论的特征选择是在一定准则下，从 $n$ 个特征中选出 $k$ 个来反映原有模式。
- 这种简单删掉某 $n-k$ 个特征的做法并不十分理想，因为一般来说，原来的 $n$ 个数据各自在不同程度上反映了识别对象的某些特征，简单地删去某些特征可能会丢失较多的有用信息。
- 如果将原来的特征做正交变换，获得的每个数据都是原来 $n$ 个数据的线性组合，然后从新的数据中选出少数几个，使其尽可能多地反映各类模式之间的差异，而这些特征间又尽可能相互独立，则比单纯的选择方法更灵活、更有效。
- K-L变换就是一种适用于任意概率密度函数的正交变换。

## 4.3 离散K-L变换

### 4.3.1 离散的有限K-L展开

- 展开式的形式

- 如果对 $c$ 种模式类别 $\{\omega_i\}_{i=1,\dots,c}$ 做离散正交展开, 则对每一模式可分别写成:  $\mathbf{x}_i = \Phi \mathbf{a}_i$ , 其中矩阵 $\Phi$ 取决于所选用的正交函数。
- 对各个模式类别, 正交函数都是相同的, 但其展开系数向量 $\mathbf{a}_i$ 则因类别的不同模式分布而异。

- K-L展开式的性质

- K-L展开式的根本性质是将随机向量 $\mathbf{x}$ 展开为另一组正交向量 $\phi_j$ 的线性和, 且其展开式系数 $\mathbf{a}_j$  (即系数向量 $\mathbf{a}$ 的各个分量) 具有不同的性质。
- 在此条件下, 正交向量集 $\{\phi_j\}$ 的确定
- K-L展开式系数的计算步骤

## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

- K-L展开式用于特征选择相当于一种线性变换。
- 若从n个特征向量中取出m个组成变换矩阵 $\Phi$ ，即

$$\Phi = (\varphi_1 \varphi_2 \dots \varphi_m), \quad m < n$$

此时， $\Phi$ 是一个 $n \times m$ 维矩阵， $x$ 是 $n$ 维向量，经过 $\Phi^T x$ 变换，即得到降维为 $m$ 的新向量。

- 问题：选取变换矩阵 $\Phi$ ，使得降维后的新向量在最小均方差条件下接近原来的向量 $x$

## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

- 结论

- 从K-L展开式的性质和按最小均方差的准则来选择特征，应使 $E[a_j]=0$ 。由于 $E[a]=E[\Phi^T x]=\Phi^T E[x]$ ，故应使 $E[x]=0$ 。基于这一条件，在将整体模式进行K-L变换之前，应先将其均值作为新坐标轴的原点，采用协方差矩阵 $C$ 或自相关矩阵 $R$ 来计算特征值。如果 $E[x] \neq 0$ ，则只能得到“次最佳”的结果。

## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

- 结论

- 将K-L展开式系数 $a_j$ （亦即变换后的特征）用 $y_j$ 表示，写成向量形式： $y = \Phi^T x$ 。此时变换矩阵 $\Phi$ 用 $m$ 个特征向量组成。为使误差最小，不采用的特征向量，其对应的特征值应尽可能小。因此，将特征值按大小次序标号，即

$$\lambda_1 > \lambda_2 > \dots > \lambda_m > \dots > \lambda_n \geq 0$$

若首先采用前面的 $m$ 个特征向量，便可使变换误差最小。此时的变换矩阵为

$$\Phi^T = \begin{pmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_m^T \end{pmatrix}$$

## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

- 结论

- K-L变换是在均方误差最小的意义下获得数据压缩（降维）的最佳变换，且不受模式分布的限制。对于一种类别的模式特征提取，它不存在特征分类问题，只是实现用低维的 $m$ 个特征来表示原来高维的 $n$ 个特征，使其误差最小，亦即使其整个模式分布结构尽可能保持不变。



## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

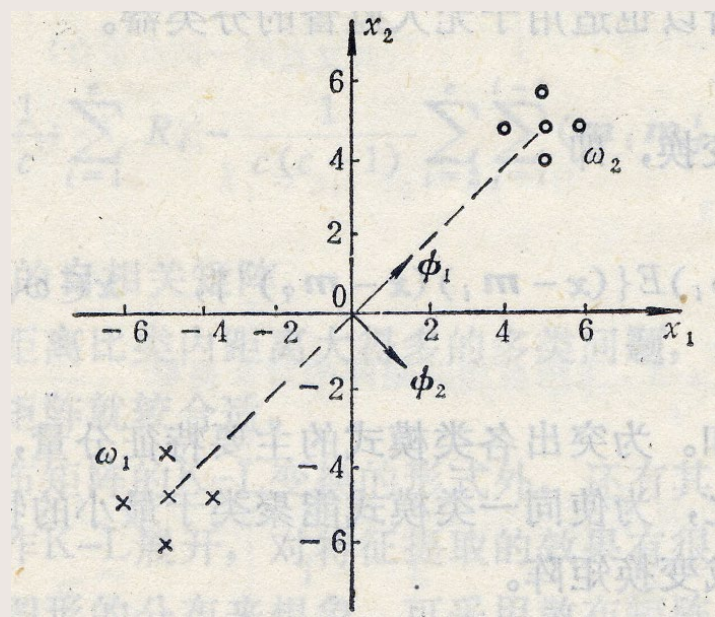
- 结论

- 通过K-L变换能获得互不相关的新特征。若采用较大特征值对应的特征向量组成变换矩阵，则能对应地保留原模式中方差最大的特征成分，所以K-L变换起到了减小相关性、突出差异性的效果。在此情况下，K-L变换也称为**主成分变换（PCA变换）**。
- 需要指出的是，采用K-L变换作为模式分类的特征提取时，要特别注意保留不同类别的模式分类鉴别信息，仅单纯考虑尽可能代表原来模式的主成分，有时并不一定有利于分类的鉴别。

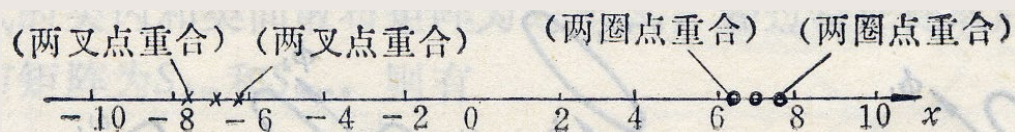
## 4.3 离散K-L变换

### 4.3.2 按K-L展开式选择特征

- [K-L变换实例]
  - 原始模式分布



– 特征提取



# 作业

- 设有如下两类样本集，其出现的概率相等：

$$\omega_1: \{(0 \ 0 \ 0)^T, (2 \ 0 \ 0)^T, \\ (2 \ 0 \ 1)^T, (1 \ 2 \ 0)^T\}$$

$$\omega_2: \{(0 \ 0 \ 1)^T, (0 \ 1 \ 0)^T, \\ (0 \ -2 \ 1)^T, (1 \ 1 \ -2)^T\}$$

用K-L变换，分别把特征空间维数降到二维和一维，并画出样本在该空间中的位置（可用matlab计算）。