第9讲 特征选择与提取1
——特征选择

周文晖

杭州电子科技大学

- **特征选择与提取基本概念**
  基本概念，特征形成，特征评价准则…
- **类别可分性测度**
  基于距离、概率分布、熵函数的可分判据 …
- **特征选择的最优搜索方法**
  穷举，分支定界，遗传算法 …
- **特征提取方法**
  基于类别可分离性判据的特征提取 …
- **主成分分析**
  PCA基本原理 …

特征选择与提取基本概念

基本概念，特征形成，特征评价准则…

类别可分性测度

基于距离、概率分布、熵函数的可分判据 …

特征选择的最优搜索方法

穷举，分支定界，遗传算法 …

特征提取方法

基于类别可分离性判据的特征提取 …

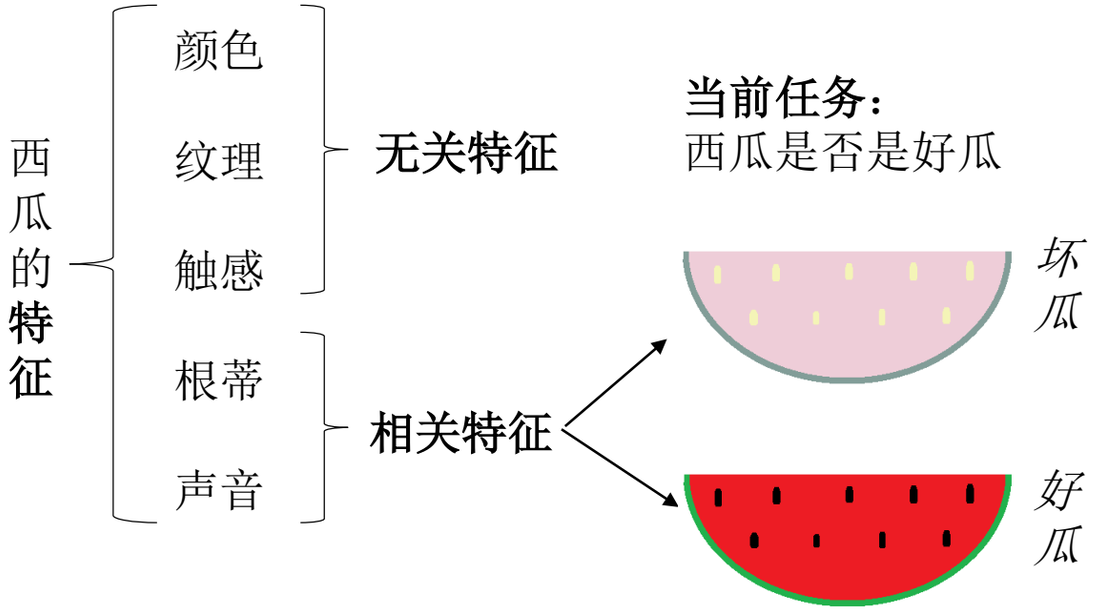
主成分分析

PCA基本原理 …

基本概念

- 特征
  - 描述物体的属性

- 特征的分类
  - 相关特征：对**当前学习任务**有用的属性
  - 无关特征：与**当前学习任务**无关的属性
  - 冗余特征：其所包含信息能由其他特征推演出来



基本概念

提出实际问题：

① 从可实现性或成本上考虑，所获得的测量值为数不多。

② 能获得的性质测量值很多。若直接作为分类特征，费时费力，且分类效果不一定好。即：“特征维数灾难”。

特征选择和提取的目的：

① 经过选择或变换，选出**任务相关**特征子集，组成识别特征，尽可能保留分类信息；

② 在保证一定分类精度的前提下，减少特征维数，使分类器的工作即快又准确。

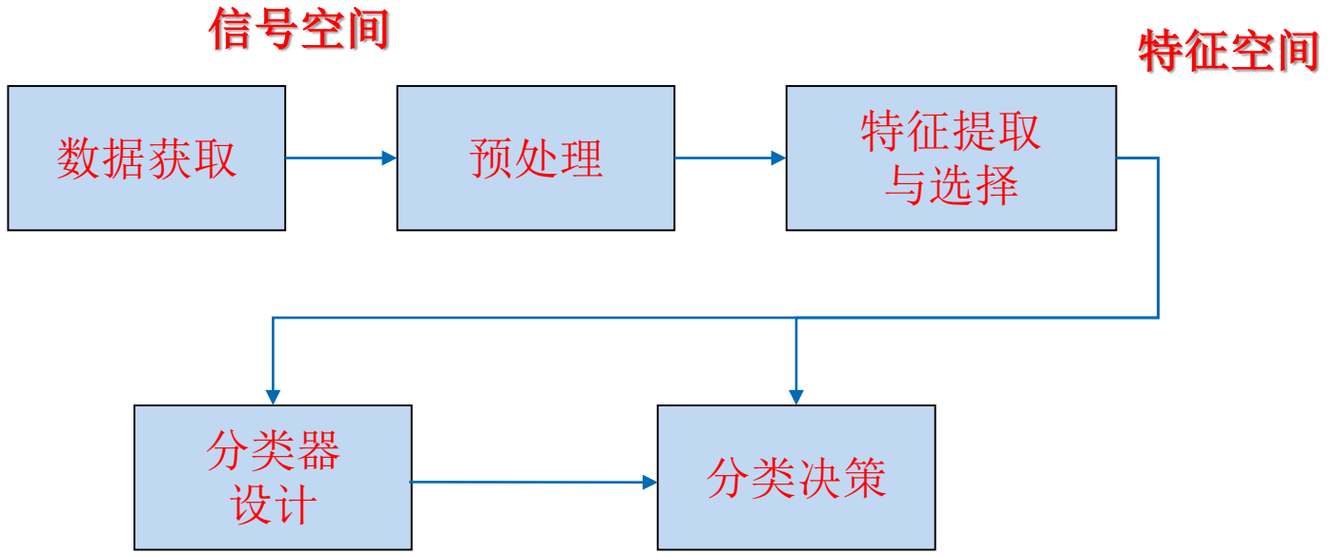
③ 减轻维度灾难：在少量属性上构建模型

④ 降低学习难度：留下关键信息

基本概念

特征的选择与提取是模式识别中重要而困难的一个环节：

- 分析各种特征的有效性并选出最有代表性的特征是模式识别的关键一步。
- 降低特征维数在很多情况下是有效设计分类器的重要课题。



基本概念

特征的要求： (1) 具有很大的识别信息量。即应具有很好的可分性。

(2) 具有可靠性。模棱两可、似是而非、时是时非等不易判别的特征应丢掉。

(3) 尽可能强的独立性。重复的、相关性强的特征只选一个。

(4) 数量尽量少，同时损失的信息尽量小。

特征选择和特征提取的异同

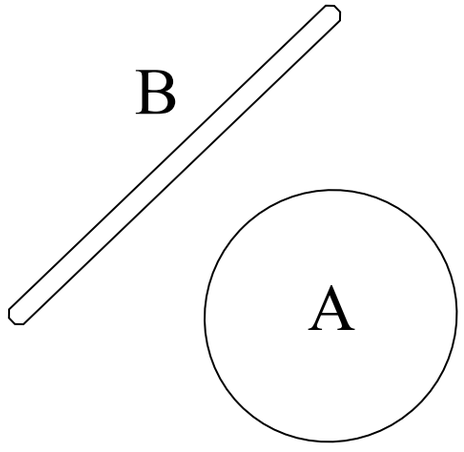
（1）特征选择：从L个度量值集合$\{x_1, x_2, \ldots, x_L\}$中按一定准则选出供分类用的子集，作为降维（$m$维，$m < L$）的分类特征。

（2）特征提取：使一组度量值$(x_1, x_2, \ldots, x_L)$通过某种变换$h_i(\cdot)$产生新的m个特征$(y_1, y_2, \ldots, y_m)$，作为降维的分类特征，其中$i = 1, 2, \ldots, m;\ m < L$。

当模式在空间中发生移动、旋转、缩放时，特征值应保持不变，保证仍可得到同样的识别效果。

基本概念

例：对一个条形和圆进行识别



解：[法1]

① 特征抽取：测量三个结构特征

(a) 周长

(b) 面积

(c) 两个互相垂直的内径比

② 分析：

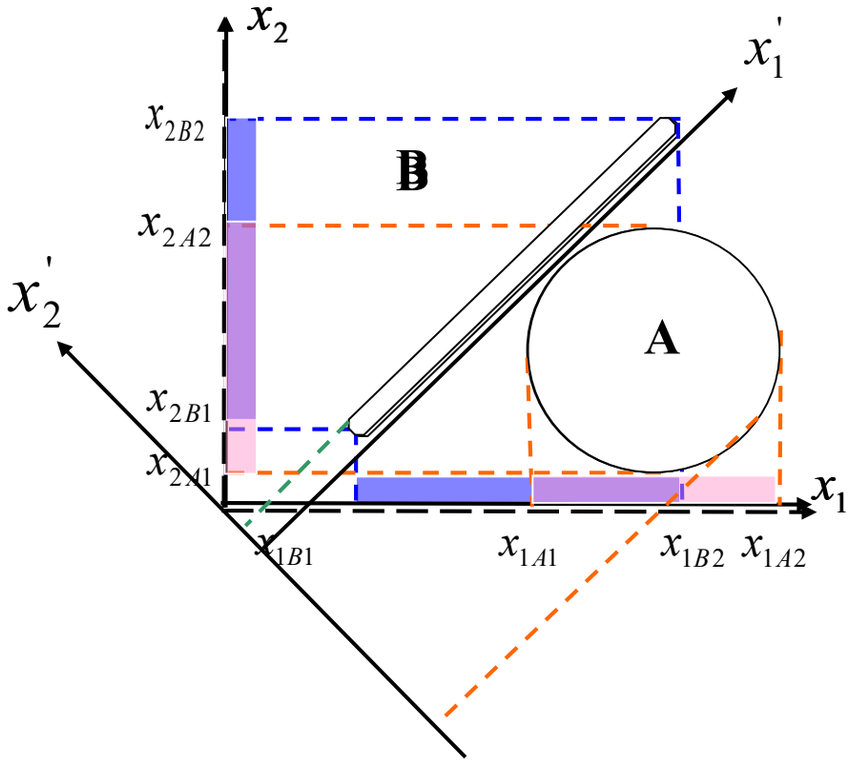
(c)是具有分类能力的特征，故选(c)，

扔掉(a) 、 (b) 。

—— 特征选择：一般根据物理特征或结构特征进行压缩。

基本概念

例：对一个条形和圆进行识别



[法2]：① 特征抽取：测量物体向两个坐标轴的投影值，则A、B各有2个值域区间。可以看出，两个物体的投影有重叠，直接使用投影值无法将两者区分开。

② 分析：将坐标系按逆时针方向做一旋转变化，或物体按顺时针方向变，并适当平移等。根据物体在 x_2' 轴上投影的坐标值的正负可区分两个物体。

——特征提取，一般用数学的方法进行压缩。

基本概念

三大类特征：物理、结构和数学特征

物理和结构特征：易于人的直觉感知，但有时难于定量描述，因而不易用于机器判别。

数学特征：易于用机器定量描述和判别，如基于统计的特征。

例：两类鱼

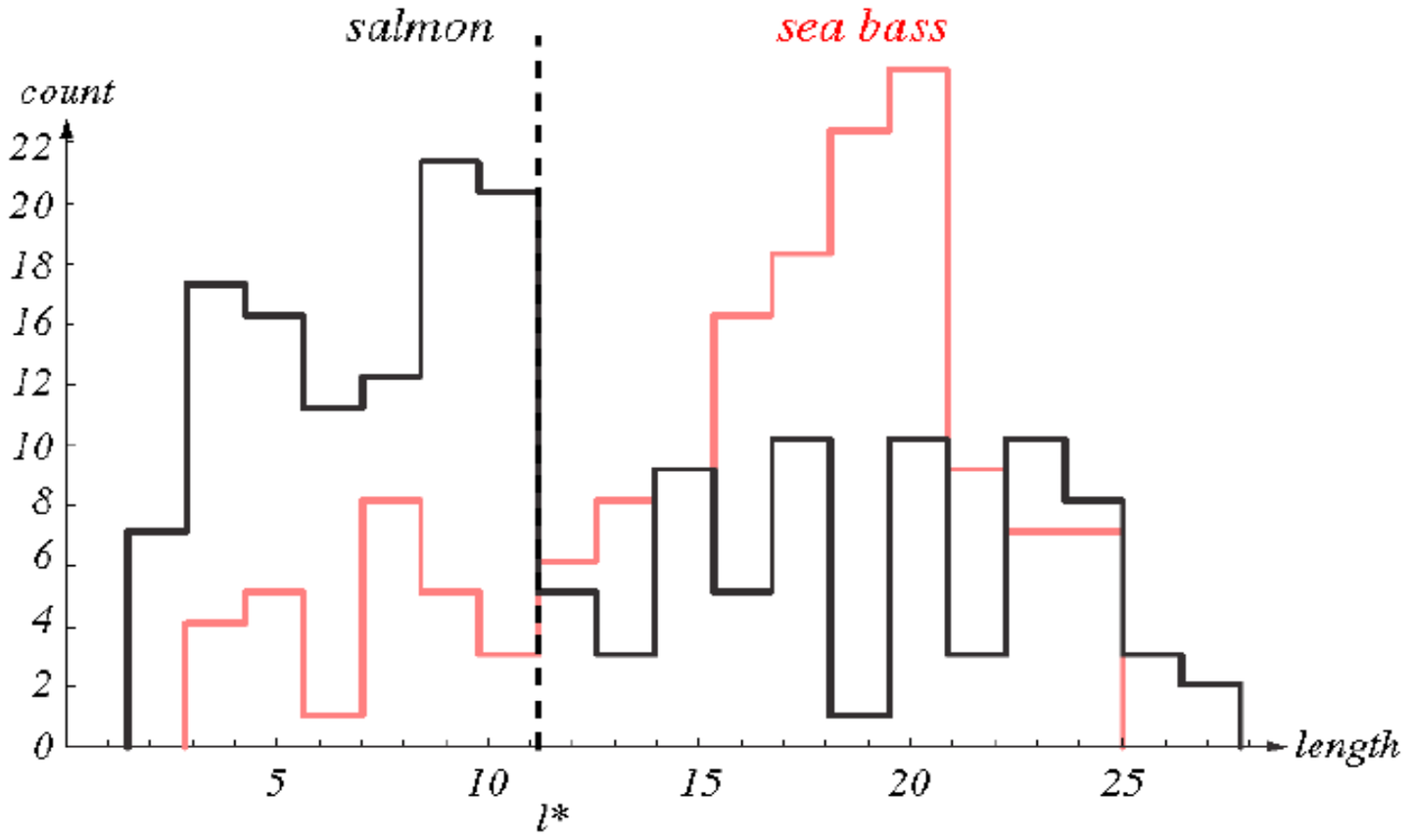
- Sea bass （海鲈鱼）
- Salmon （三文鱼）

Pattern Classification, 2001

基本概念

三大类特征

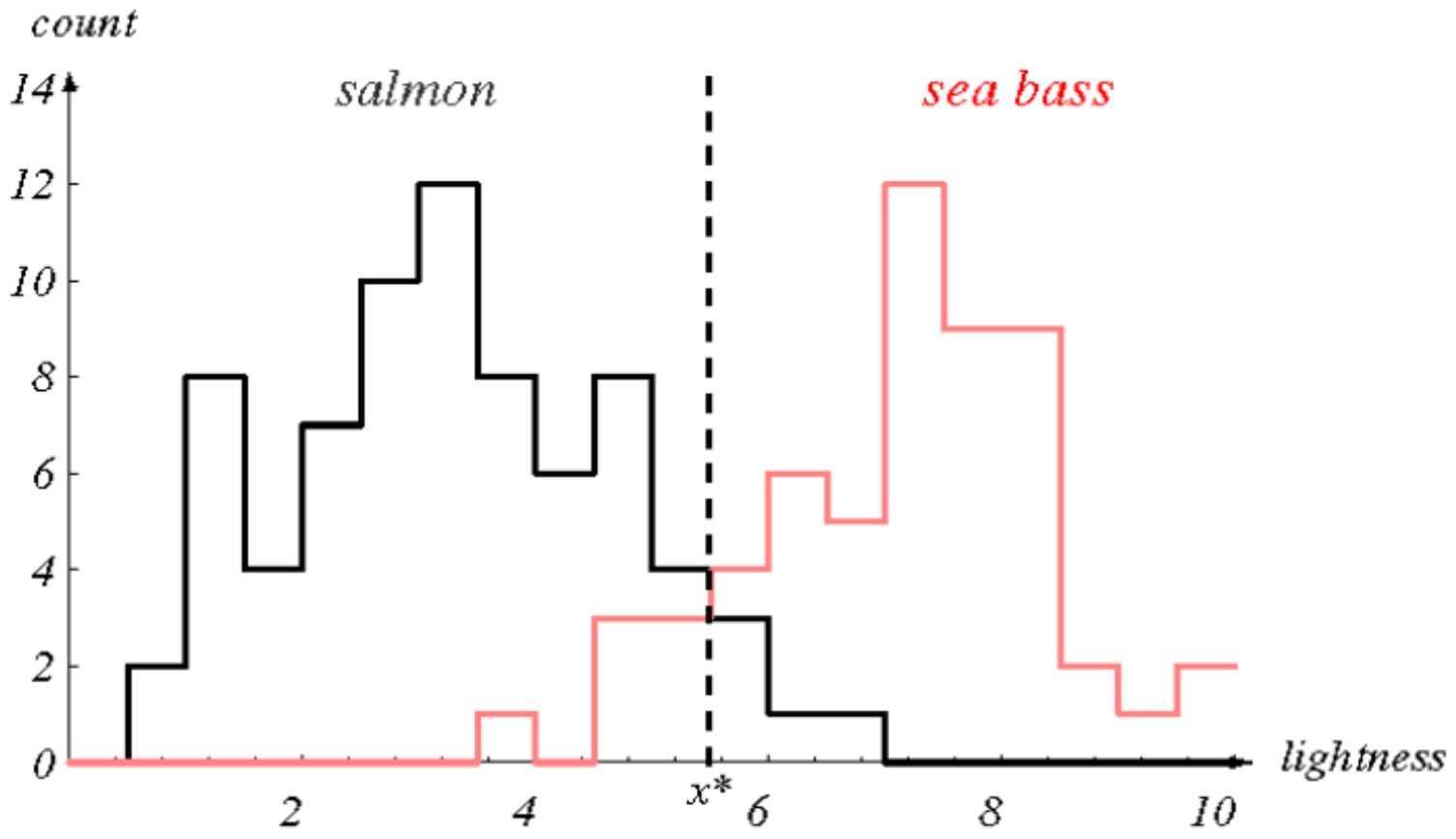
特征1：长度



基本概念

三大类特征

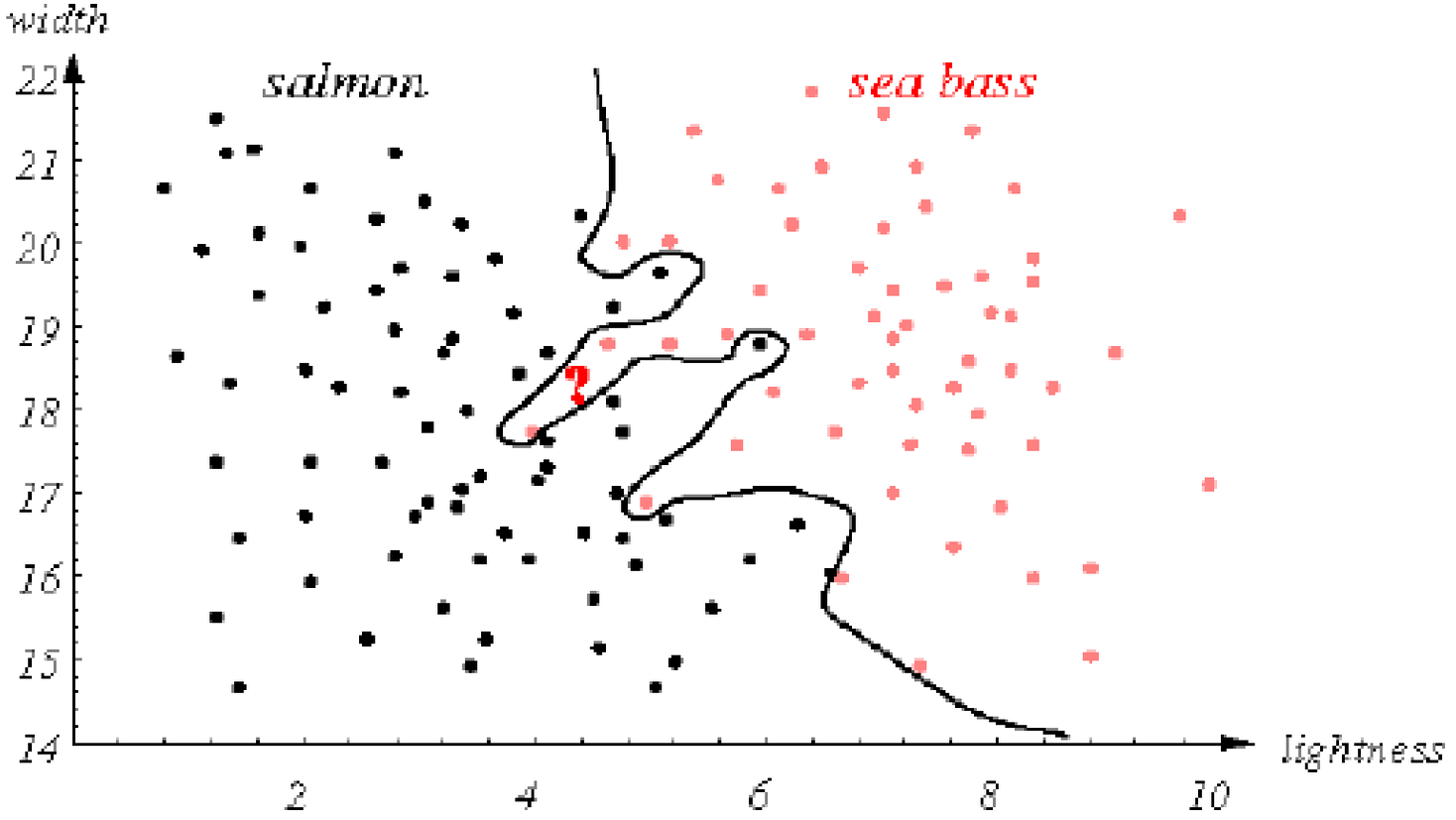
特征2：亮度



基本概念

三大类特征

模式分类：线性、二次、最近邻分类器



基本概念

特征的形成

- **特征形成** (acquisition)：
  - 信号获取或测量→原始测量
  - 原始特征
- 实例：
  - 数字图象中的各像素灰度值
  - 人体的各种生理指标
- 原始特征分析：
  - 原始测量不能反映对象本质
  - 高维原始特征不利于分类器设计：计算量大，冗余，样本分布十分稀疏。

基本概念

特征的形成

- 两类提取有效信息、压缩特征空间的方法：特征提取和特征选择
- **特征提取** (extraction)：用映射（或变换）的方法把原始特征变换为较少的新特征。
- **特征选择**(selection)：从原始特征中挑选出一些最有代表性，分类性能最好的特征。
- 特征的选择与提取与具体问题有很大关系，目前没有理论能给出对任何问题都有效的特征选择与提取方法。

基本概念

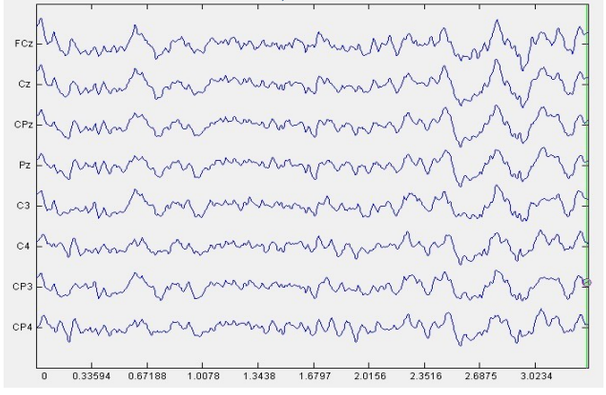
特征的形成

例 细胞自动识别：

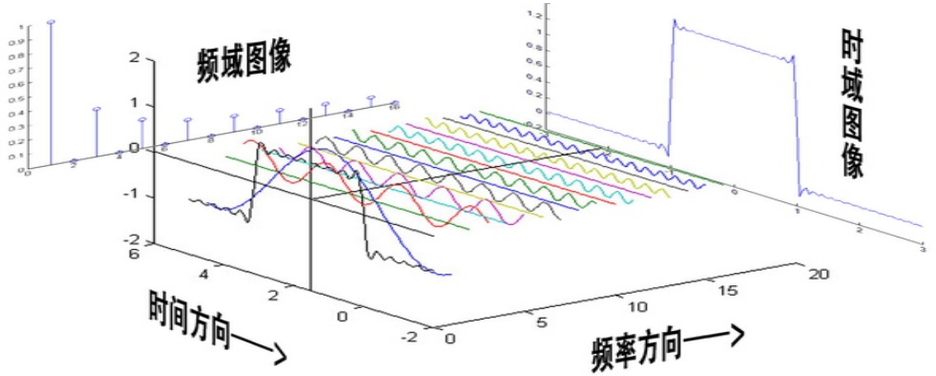
- **原始测量**：（正常与异常）细胞的数字图像
- **原始特征**（特征的形成，找到一组代表细胞性质的特征）：细胞面积，胞核面积，形状系数，光密度，核内纹理，核浆比
- **压缩特征**：原始特征的维数仍很高，需压缩以便于分类
  - 特征选择：挑选最有分类信息的特征：专家知识，数学方法
  - 特征提取：数学变换
    - 傅立叶变换或小波变换
    - 用PCA方法作特征压缩

基本概念

特征的形成

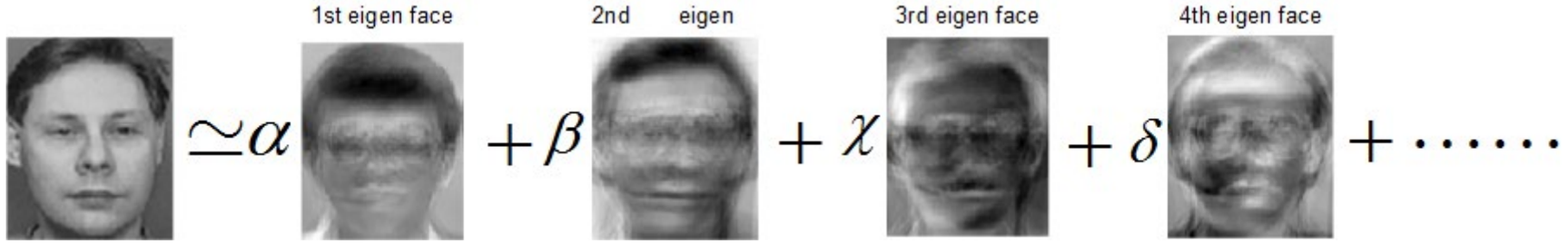


特征提取之傅里叶变化（抽象）

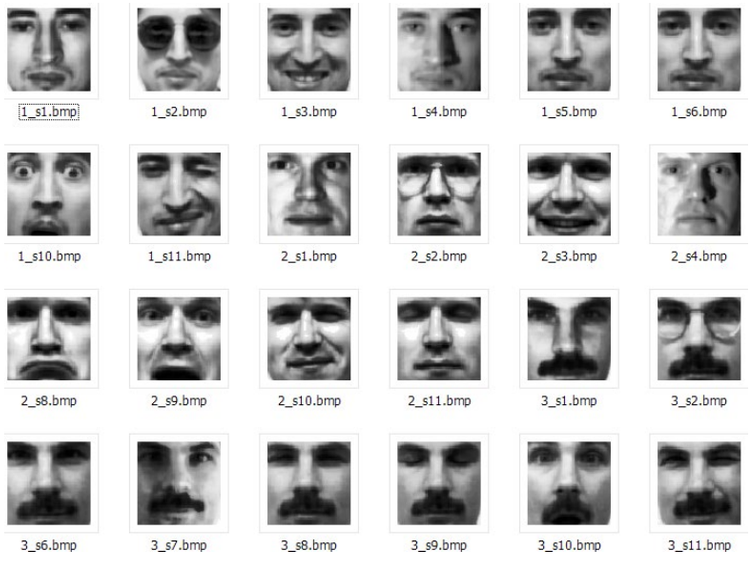


基本概念

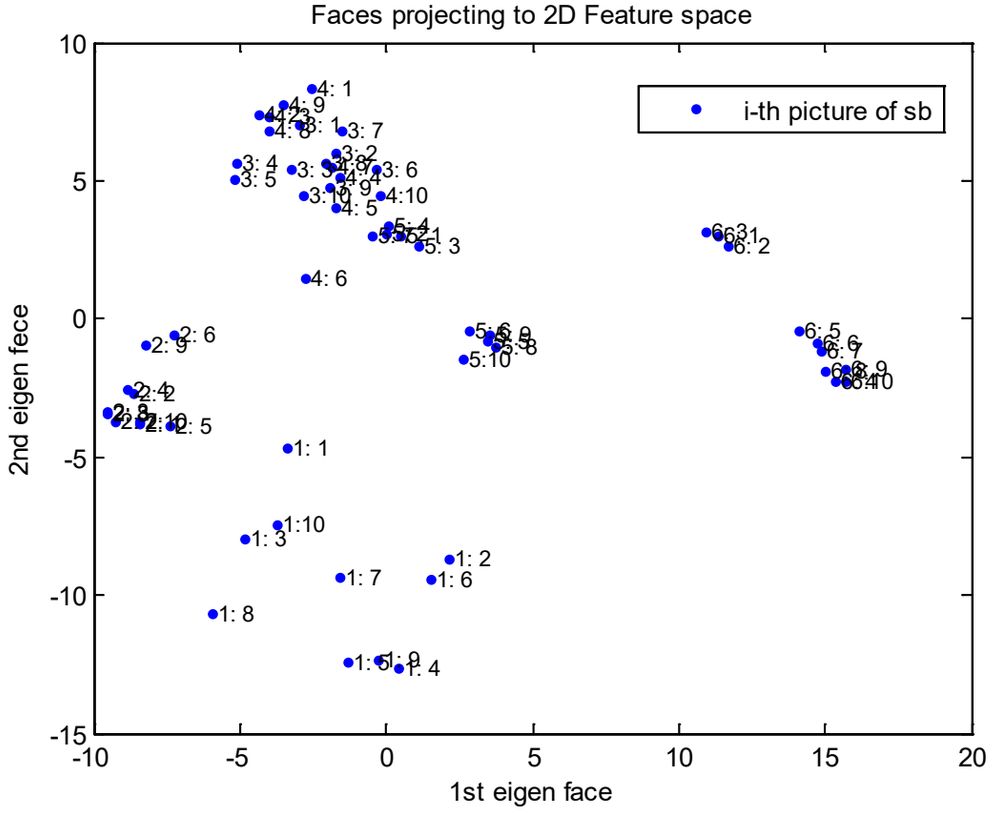
特征的形成



人脸识别中的特征提取（本征脸）



人脸库部分样本



- **特征选择与提取基本概念**
  基本概念，特征形成，特征评价准则…
- **类别可分性测度**
  基于距离、概率分布、熵函数的可分判据 …
- **特征选择的最优搜索方法**
  穷举，分支定界，遗传算法 …
- **特征提取方法**
  基于类别可分离性判据的特征提取 …
- **主成分分析**
  PCA基本原理 …

基本概念

特征的评价准则

特征选择与提取的任务是找出一组对分类最好的特征 ➔ 评价准则

从 D 个原始特征中选出或转换成 $d\,(<D)$ 个特征，有很多种组合，哪种组合好需要有个比较标准。

概念：

数学上定义的用以衡量特征对分类的效果的准则

实际问题中需根据实际情况人为确定

- **误识率判据：**

  理论上的目标，实际采用困难（密度未知，形式复杂，样本不充分，…）

- **可分性判据**：实用的、可计算的判据，即**类别可分性测度。**

类别可分性测度

类别可分性测度：衡量不同特征及其组合对分类是否有效的定量准则。

相似性测度：衡量模式之间相似性的一种尺度

令 J_{ij} 是第 i 类和第 j 类的可分性测度函数， J_{ij} 愈大，则两类的分离程度就愈大。

性质：

$J_{ij} > 0$ 当 $i \neq j$ 时

$J_{ij} = 0$ 当 $i = j$ 时

$J_{ij} = J_{ji}$

单调性：增加新特征，判据不减小

$$J_{ij}\left(x_1, x_2, ..., x_d\right) \leq J_{ij}\left(x_1, x_2, ..., x_d, x_{d+1}\right)$$

类别可分性测度

- 空间分布： 类内距离和类间距离
- 随机模式向量： 类概率密度函数
- 错误率 → 与错误率有关的距离

- 常见类别可分离性判据：基于距离、概率分布、熵函数

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度　　思想：类间可分性 = 所有样本间的平均距离。

1．类内距离和类内散布矩阵

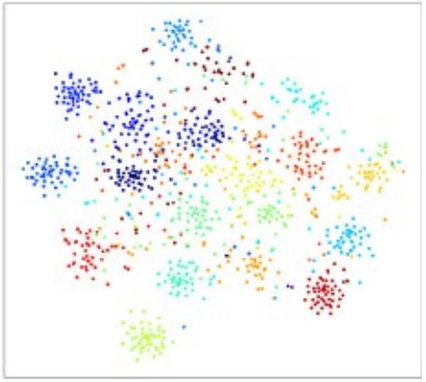
1) 类内距离：同一类模式点集内，各样本间的均方距离。

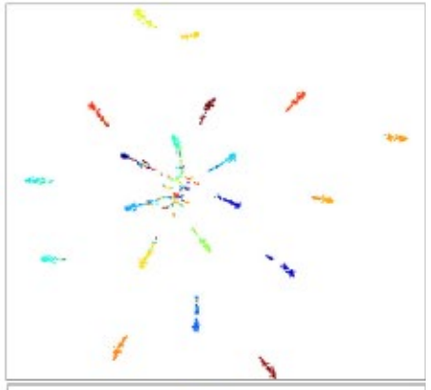
平方形式：$\bar{D}^2 = E\left\{\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|^2\right\} = E\left\{\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)^{\mathrm{T}}\left(\boldsymbol{X}_i - \boldsymbol{X}_j\right)\right\}$

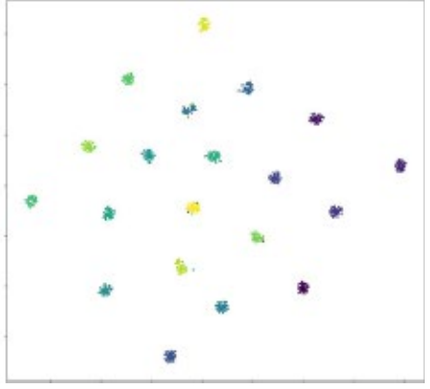
$\boldsymbol{X}_i,\ \boldsymbol{X}_j$：$n$ 维模式点集$\{\boldsymbol{X}\}$中的任意两个样本 。

若$\{\boldsymbol{X}\}$中的样本相互独立，有 $\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_j = \boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_i = 0$ ，

$\bar{D}^2 = 2E\left\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right\} - 2E\left\{\boldsymbol{X}^{\mathrm{T}}\right\}E\{\boldsymbol{X}\} = 2\left[E\left\{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right\} - \boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\right]$





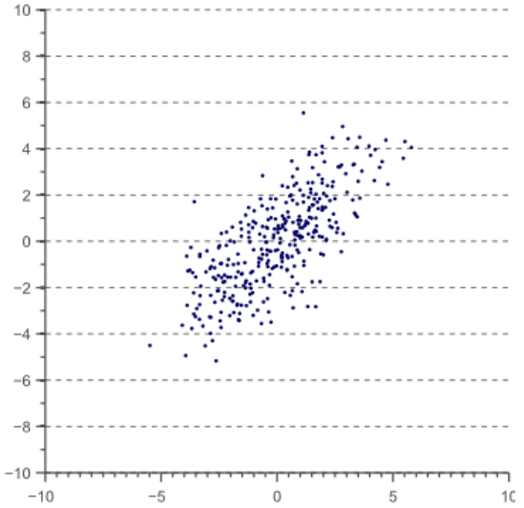


类别可分性测度——基于距离的可分性测度

基于距离的可分性测度　　思想：类间可分性 = 所有样本间的平均距离。

1．类内距离和类内散布矩阵

2) 类内散布矩阵：表示各样本点围绕均值的散布情况 ——该类分布的协方差矩阵。



协方差的几何意义：

- ✓ 方差只能说明数据在特征空间坐标轴上的离散度。如x轴上的方差和y轴上的方差；
- ✓ 但数据在水平和垂直方向上的离散度，无法解释数据呈对角线分布的特点。
- ✓ 为表示数据x值和y值间的相关性，将方差概念推广至“协方差”：

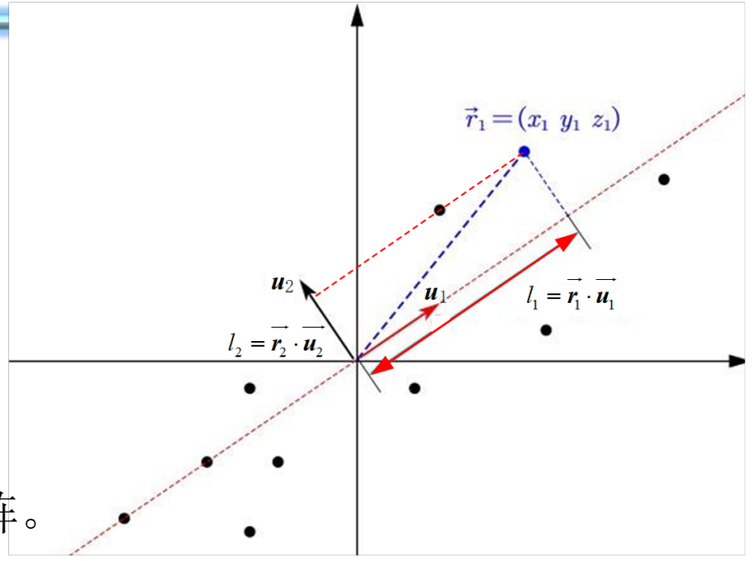
$$\sigma(x,y)=E\left[\left(x-E(x)\right)\left(y-E(y)\right)\right]$$

- ✓ 对于二维数据，可得到$\sigma(x,x)$, $\sigma(x,y)$, $\sigma(y,x)$, $\sigma(y,y)$，构成协方差矩阵：

$$\Sigma=\begin{bmatrix}\sigma(x,x) & \sigma(x,y)\\ \sigma(y,x) & \sigma(y,y)\end{bmatrix}$$

由于x与y的相关性，等价于y与x的相关性。因此，协方差矩阵是一个由位于对角线上的方差和非对角线上的协方差构成的对称矩阵。

类别可分性测度——基于距离的可分性测度

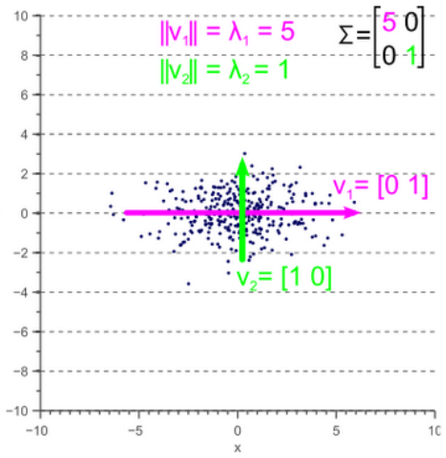


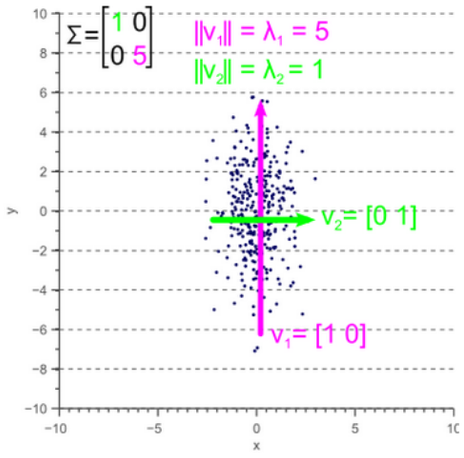
基于距离的可分性测度　　思想：类间可分性 = 所有样本间的平均距离。

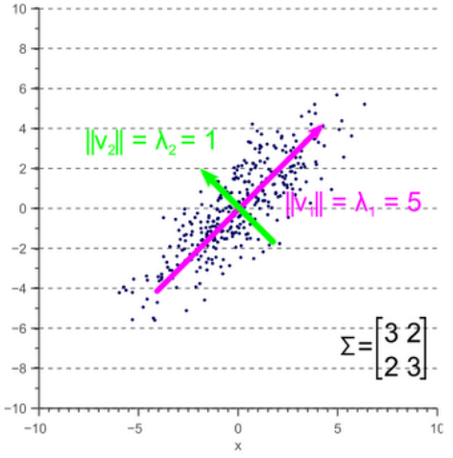
1．类内距离和类内散布矩阵

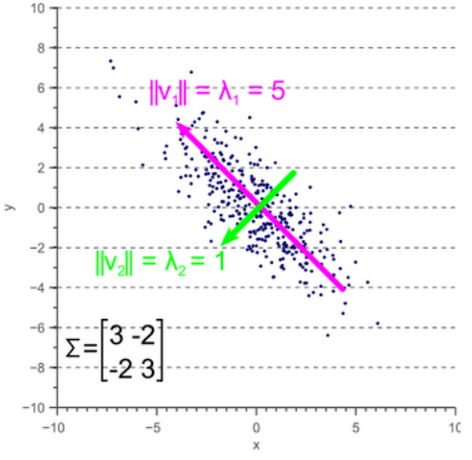
2) 类内散布矩阵：表示各样本点围绕均值的散布情况 ——该类分布的协方差矩阵。

协方差矩阵的特征值分解：

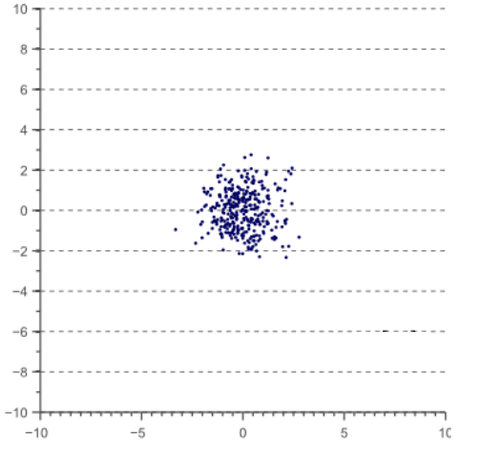








优选



协方差矩阵为对角矩阵，即协方差部分为零，则方差等于特征值 λ 。

协方差矩阵为非对角矩阵，特征值是特征向量上的方差。

特征选择和提取的结果应使类内散布矩阵的迹愈小愈好。

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

类间可分性 = 所有样本间的平均距离：

2．类间距离和类间散布矩阵

1) 类间距离：模式类之间的距离，记为$\overline{D_b}$。

$$\overline{D_b^2}=\sum_{i=1}^{c}P(\omega_i)\,\|\boldsymbol{M}_i-\boldsymbol{M}_0\|^2=\sum_{i=1}^{c}P(\omega_i)(\boldsymbol{M}_i-\boldsymbol{M}_0)^{\mathrm{T}}(\boldsymbol{M}_i-\boldsymbol{M}_0)$$

$$\boldsymbol{M}_0=E\{\boldsymbol{X}\}=\sum_{i=1}^{c}P(\omega_i)\boldsymbol{M}_i \qquad \boldsymbol{X}\in\omega_i,\ i=1,2,...,c$$

每类模式均值向量与模式总体均值向量之间平方距离的先验概率加权和。

式中，$P(\omega_i)$：ω_i类的先验概率；

$\boldsymbol{M}_i$：ω_i类的均值向量；

$\boldsymbol{M}_0$：所有 c 类模式的总体均值向量。

2) 类间散布矩阵：表示c类模式在空间的散布情况，记为$\boldsymbol{S}_b$。 $\boldsymbol{S}_b=\sum_{i=1}^{c}P(\omega_i)(\boldsymbol{M}_i-\boldsymbol{M}_0)(\boldsymbol{M}_i-\boldsymbol{M}_0)^{\mathrm{T}}$

注意：与类间距离的转置位置不同。

3) 类间距离与类间散布矩阵的关系： $\overline{D_b^2}=\mathrm{tr}\{\boldsymbol{S}_b\}$

类间散布矩阵的迹愈大愈有利于分类。

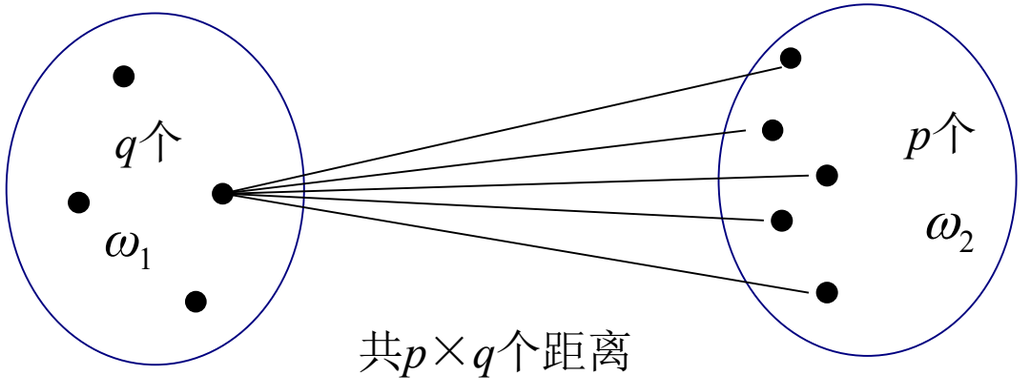
类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

3．多类模式向量间的距离和总体散布矩阵

1）两类情况的距离

设 ω_1 类中有 q 个样本， ω_2 类中有 p 个样本。



两个类区之间的距离 = $p \times q$个距离的平均距离

类似地 多类情况

多类间任意两个点间距离的平均距离

多类间任意两个点间**平方距离**的平均值

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

3．多类模式向量间(各类间)的距离和总体散布矩阵

2）多类情况的距离

任意类的组合

特定两类间任意样本的组合

多类模式向量间的平均平方距离 J_d

$$J_d = \frac{1}{2}\sum_{i=1}^{c} P(\omega_i)\sum_{j=1}^{c} P(\omega_j)\frac{1}{n_i n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} D^2(\boldsymbol{X}_k^i, \boldsymbol{X}_l^j)$$

式中，$P(\omega_i)$ 和 $P(\omega_j)$：ω_i 和 ω_j 类先验概率；c：类别数；

$\boldsymbol{X}_k^i$：ω_i 类的第 k 个样本；$\boldsymbol{X}_l^j$：ω_j 类的第 l 个样本；

n_i 和 n_j：ω_i 和 ω_j 类的样本数；

$D^2(\boldsymbol{X}_k^i, \boldsymbol{X}_l^j)$：$\boldsymbol{X}_k^i$ 和 $\boldsymbol{X}_l^j$ 间欧氏距离的平方。

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

3．多类模式向量间的距离和总体散布矩阵

2）平均平方距离 J_d 的另一种形式

平方距离：$D^2(\boldsymbol{X}_k^i,\boldsymbol{X}_l^j)=(\boldsymbol{X}_k^i-\boldsymbol{X}_l^j)^{\mathrm{T}}(\boldsymbol{X}_k^i-\boldsymbol{X}_l^j)$

ω_i 类的均值向量：$\boldsymbol{M}_i=\dfrac{1}{n_i}\sum_{k=1}^{n_i}\boldsymbol{X}_k^i$

c类模式总体的均值向量：$\boldsymbol{M}_0=\sum_{i=1}^{c}P(\omega_i)\boldsymbol{M}_i$

代入 J_d 公式

$$J_d=\frac{1}{2}\sum_{i=1}^{c}P(\omega_i)\sum_{j=1}^{c}P(\omega_j)\frac{1}{n_in_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j}D^2(\boldsymbol{X}_k^i,\boldsymbol{X}_l^j)$$

得:

$$J_d=\sum_{i=1}^{c}P(\omega_i)\left[\frac{1}{n_i}\sum_{k=1}^{n_i}(\boldsymbol{X}_k^i-\boldsymbol{M}_i)^{\mathrm{T}}(\boldsymbol{X}_k^i-\boldsymbol{M}_i)+(\boldsymbol{M}_i-\boldsymbol{M}_0)^{\mathrm{T}}(\boldsymbol{M}_i-\boldsymbol{M}_0)\right]$$

某类类内平方距离平均值

某类类间平方距离

+

某类的平方距离

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

3．多类模式向量间的距离和总体散布矩阵

3）多类情况的散布矩阵

多类类间散布矩阵：$\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{M}_i - \boldsymbol{M}_0)(\boldsymbol{M}_i - \boldsymbol{M}_0)^{\mathrm{T}}$

多类类内散布矩阵：$\boldsymbol{S}_w = \sum_{i=1}^{c} P(\omega_i) E\{(\boldsymbol{X} - \boldsymbol{M}_i)(\boldsymbol{X} - \boldsymbol{M}_i)^{\mathrm{T}}\} \qquad \boldsymbol{X} \in \omega_i$

$$= \sum_{i=1}^{c} P(\omega_i) \frac{1}{n_i} \sum_{k=1}^{n_i} (\boldsymbol{X}_k^i - \boldsymbol{M}_i)(\boldsymbol{X}_k^i - \boldsymbol{M}_i)^{\mathrm{T}}$$

—— 各类模式协方差矩阵的先验概率加权平均值。

多类模式的总体散布矩阵：

$$\boldsymbol{S}_t = E\{(\boldsymbol{X} - \boldsymbol{M}_0)(\boldsymbol{X} - \boldsymbol{M}_0)^{\mathrm{T}}\} = \boldsymbol{S}_b + \boldsymbol{S}_w$$

类别可分性测度——基于距离的可分性测度

基于距离的可分性测度

3．多类模式向量间的距离和总体散布矩阵

4）多类模式平均平方距离与总体散布矩阵的关系

$$J_d = \mathrm{tr}(\boldsymbol{S}_t) = \mathrm{tr}(\boldsymbol{S}_b + \boldsymbol{S}_w)$$

tr：矩阵的迹（方阵主对角线上各元素之和）。

常用的基于类内类间距离的可分性判据

$$J_1 = \mathrm{tr}\left(\mathrm{S}_{\mathrm{w}} + S_b\right) \qquad J_4 = \frac{\mathrm{tr}S_b}{\mathrm{tr}S_w}$$

$$J_2 = \mathrm{tr}\left(S_w^{-1} S_b\right)$$

$$J_3 = \ln\frac{|S_b|}{|S_w|} \qquad J_5 = \frac{|S_b - S_w|}{|S_w|}$$

距离与散布矩阵作为可分性测度的特点：

*** 计算方便，概念直观（反映模式的空间分布情况 ）；**

*** 与分类错误率没有直接的联系。**

类别可分性测度——基于距离的可分性测度

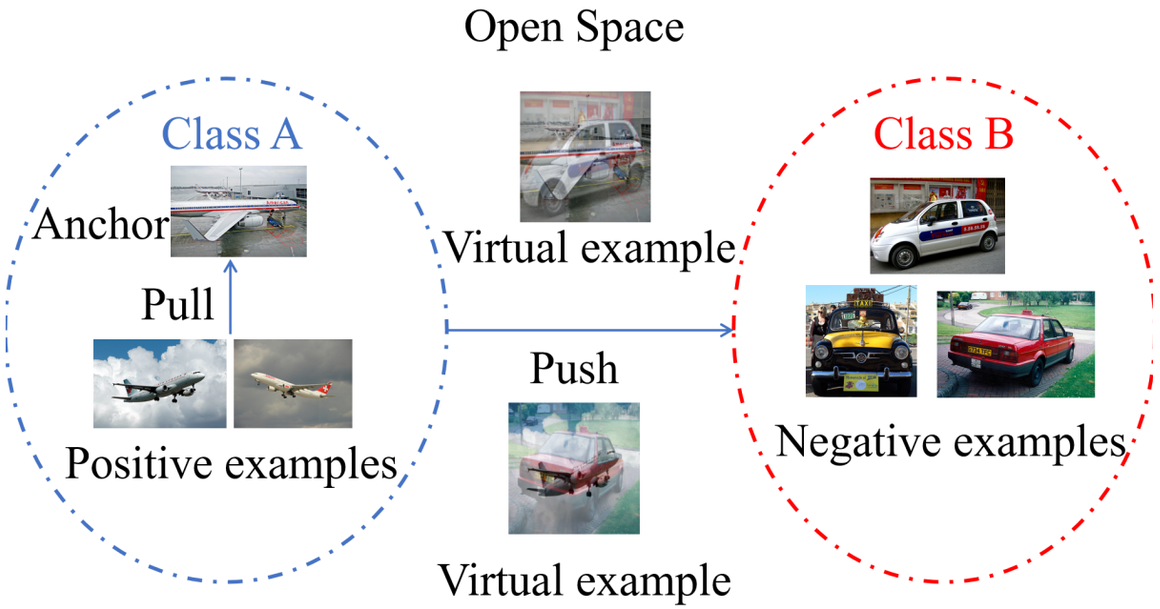


Figure 1: Overview of the proposed method. Supervised contrastive learning pulls the positive examples from the same class towards the anchor example, while pushing the negative examples away. Virtual examples generated by the mixup algorithm simulate unknown samples in the open space.

The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

DOI: https://doi.org/10.1609/aaai.v37i9.26253

Contrastive Open Set Recognition

Baile Xu[1,2], Furao Shen[1,3*], Jian Zhao[4]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Department of Computer Science and Technology, Nanjing University
[3]School of Artificial Intelligence, Nanjing University
[4]School of Electronic Science and Engineering, Nanjing University
blxu@smail.nju.edu.cn, frshen@nju.edu.cn, jianzhao@nju.edu.cn

对比学习是表征学习领域中近年来备受关注的一个研究领域。分为：

- 自我监督对比学习(Van den Oord, Li, and Vinyals 2018; He et al. 2020; Chen et al. 2020b; Chen and He 2021),
- 监督对比学习 (Khosla et al. 2020))

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

考查两类分布密度之间的交叠程度

定义：两个密度函数之间的距离

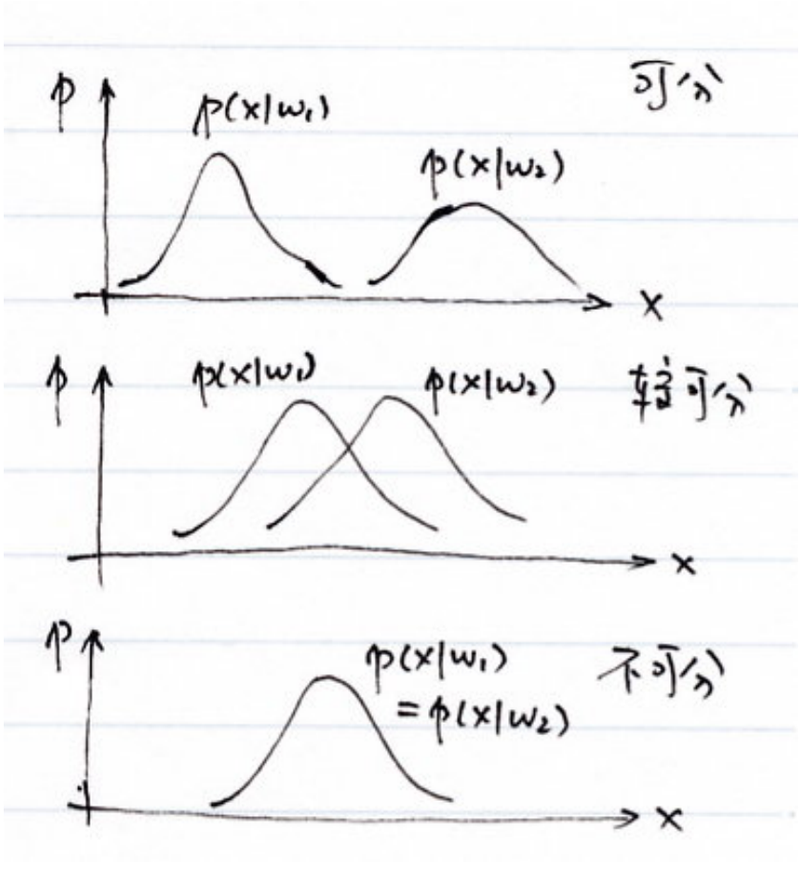
$$J_p(\cdot)=\int g\left[p(\mathbf{x}\mid\omega_1),p(\mathbf{x}\mid\omega_2),P_1,P_2\right]d\mathbf{x}$$

它必须满足三个条件：

1) $J_p\geq 0$；

2) 若 $p(x\mid\omega_1)p(x\mid\omega_2)=0,\forall x$，则 $J_p=J_{\max}$，完全不重叠；

3) 若 $p(x\mid\omega_1)=p(x\mid\omega_2),\forall x$，则 $J_p=0$，完全重叠。



类别可分性测度——基于概率分布的可分性测度

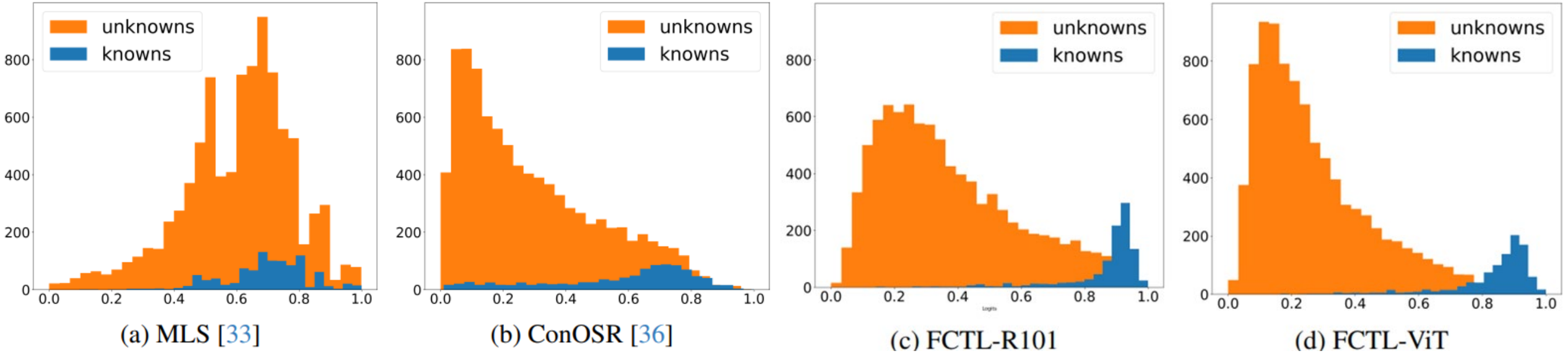


Figure s1. Visualization of the unknown-class and known-class score distributions, where the horizontal axis represents the normalized unknown scores and the vertical axis represents the number of samples. The envelopes of orange and blue areas represent the unknown-class and known-class score distribution, respectively.

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

具体距离定义有多种：

1、Bhattacharyya距离

$$J_B = -\ln \int \left[p(x \mid \omega_1) p(x \mid \omega_2) \right]^{\frac{1}{2}} dx$$

两类完全重合时，$J_B = 0$ ；两类完全不交叠时，$J_B = \infty$

2、切诺夫界（Chernoff bound）

$$J_c = -\ln \int p^s(x \mid \omega_1) p^{1-s}(x \mid \omega_2) dx$$

当s=0.5时，$J_c = J_B$

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

3、散度

1）散度的定义　　出发点：对数似然比含有类别的可分性信息。

设 ω_i, ω_j 类的概率密度函数分别为 $p(\boldsymbol{X}|\omega_i)$ 和 $p(\boldsymbol{X}|\omega_j)$

ω_i 类对 ω_j 类的对数似然比：$l_{ij} = \ln \dfrac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)}$　　ω_j 类对 ω_i 类的对数似然比：$l_{ji} = \ln \dfrac{p(\boldsymbol{X}|\omega_j)}{p(\boldsymbol{X}|\omega_i)}$

对不同的X，似然函数不同，对数似然比体现的可分性不同，通常采用平均可分性信息——对数似然比的期望值 。

ω_i 类对数似然比的期望值：$I_{ij} = E\{l_{ij}\} = \int_X p(\boldsymbol{X}|\omega_i) \ln \dfrac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)} d\boldsymbol{X}$

$E\{x\} = \int_{-\infty}^{\infty} x p(x) d(x)$

ω_j 类对数似然比的期望值：$I_{ji} = E\{l_{ji}\} = \int_X p(\boldsymbol{X}|\omega_j) \ln \dfrac{p(\boldsymbol{X}|\omega_j)}{p(\boldsymbol{X}|\omega_i)} d\boldsymbol{X}$

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

3、散度

散度等于两类的对数似然比期望值之和。

ω_i 类对 ω_j 类的散度定义为 J_{ij}：

$$J_{ij} = I_{ij} + I_{ji} = \int_X \left[p(\boldsymbol{X}|\omega_i) - p(\boldsymbol{X}|\omega_j) \right] \ln \frac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)} d\boldsymbol{X}$$

散度表示了区分ω_i类和ω_j类的总的平均信息。

——特征选择和特征提取应使散度尽可能的大。

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

2．散度的性质

（1）$J_{ij}=J_{ji}$　　$J_{ij}=I_{ij}+I_{ji}=\int_X[p(\boldsymbol{X}|\omega_i)-p(\boldsymbol{X}|\omega_j)]\ln\frac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)}d\boldsymbol{X}$

$$J_{ji}=I_{ji}+I_{ij}=\int_X[p(\boldsymbol{X}|\omega_j)-p(\boldsymbol{X}|\omega_i)]\ln\frac{p(\boldsymbol{X}|\omega_j)}{p(\boldsymbol{X}|\omega_i)}d\boldsymbol{X}$$

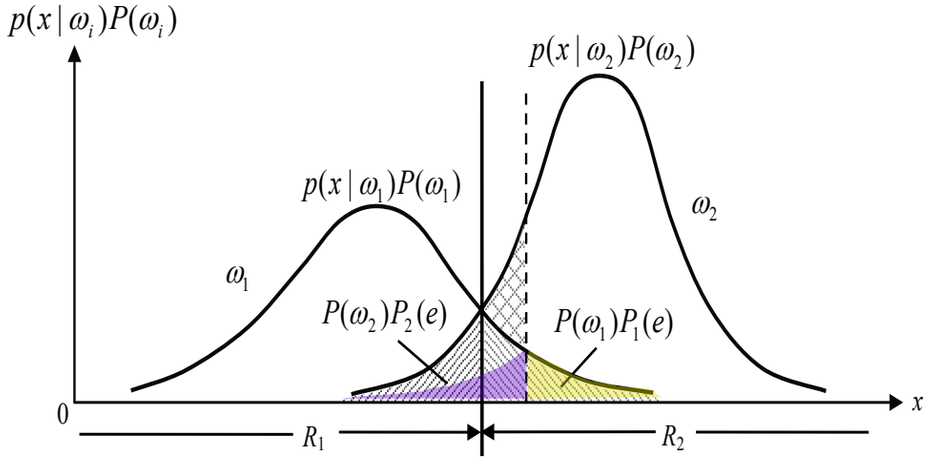
（2）J_{ij}为非负，即$J_{ij}\geq 0$。

当$p(\boldsymbol{X}|\omega_i)\neq p(\boldsymbol{X}|\omega_j)$时，$J_{ij}>0$，

$p(\boldsymbol{X}|\omega_i)$与$p(\boldsymbol{X}|\omega_j)$相差愈大，$J_{ij}$越大。

当$p(\boldsymbol{X}|\omega_i)=p(\boldsymbol{X}|\omega_j)$，两类分布密度相同，$J_{ij}=0$。

类别可分性测度——基于概率分布的可分性测度



基于概率分布的可分性测度

2．散度的性质

（3）错误率分析中，两类概率密度曲线交叠越少，错误率越小。

由散度的定义式 $J_{ij}=I_{ij}+I_{ji}=\int_X[p(\boldsymbol{X}|\omega_i)-p(\boldsymbol{X}|\omega_j)]\ln\dfrac{p(\boldsymbol{X}|\omega_i)}{p(\boldsymbol{X}|\omega_j)}d\boldsymbol{X}$

可知，散度愈大，两类概率密度函数曲线相差愈大，交叠愈少，分类错误率愈小。

（4）散度具有可加性：对于模式向量 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$，若各分量相互独立，则有

$$J_{ij}(\boldsymbol{X})=J_{ij}(x_1,x_2,...,x_n)=\sum_{k=1}^{n}J_{ij}(x_k)$$

据此可估计每一个特征在分类中的重要性：散度较大的特征含有较大的可分信息——保留。

（5）可加性表明，加入新的特征，不会使散度减小。即

$$J_{ij}(x_1,x_2,\cdots,x_n)\le J_{ij}(x_1,x_2,\cdots,x_n,x_{n+1})$$

类别可分性测度——基于概率分布的可分性测度

基于概率分布的可分性测度

3．两个正态分布模式类的散度

设ω_i类和ω_j类的概率密度函数分别为 $p(\boldsymbol{X}|\omega_i)\sim N(\boldsymbol{M}_i,\boldsymbol{C})$ 和 $p(\boldsymbol{X}|\omega_j)\sim N(\boldsymbol{M}_j,\boldsymbol{C})$

可得到 ω_i 类对 ω_j 类的散度为

$$J_{ij}=\mathrm{tr}[(\boldsymbol{C}^{-1}(\boldsymbol{M}_i-\boldsymbol{M}_j)(\boldsymbol{M}_i-\boldsymbol{M}_j)^T]=(\boldsymbol{M}_i-\boldsymbol{M}_j)^T\boldsymbol{C}^{-1}(\boldsymbol{M}_i-\boldsymbol{M}_j)$$

—— 两类模式之间马氏距离的平方

一维正态分布时：

$$J_{ij}=\frac{(m_i-m_j)^2}{\sigma^2}$$

两类均值向量距离越远，散度愈大

每类自身分布愈集中，两类间的散度愈大

类别可分性测度——基于熵的可分性测度

基于熵的可分性测度

熵：事件不确定性的度量

A事件的不确定性大（熵大），则对A事件的观察所提供的信息量大。

思路：

把各类 ω_i 看作一系列事件

把后验概率$P(\omega_i \mid x)$看作特征x上出现ω_i的概率

如从x能确定ω_i，则对ω_i的观察不提供信息量，熵为0。

—— 特征x有利于分类。

如从x完全不能确定ω_i，则对ω_i的观察信息量大，熵大。

—— 特征x无助于分类。

类别可分性测度——基于熵的可分性测度

基于熵的可分性测度

定义熵函数 $H = J_c\left[P(\omega_1 \mid x), \cdots P(\omega_c \mid x)\right]$

须满足

① 规一化 $J_c\left(\frac{1}{c}, \cdots, \frac{1}{c}\right) = 1$

$$0 \le J_c(P_1, \cdots, P_c) \le J_c\left(\frac{1}{c}, \cdots, \frac{1}{c}\right) = 1$$

② 对称性 $J_c(P_1, \cdots, P_c) = J_c(P_?, \cdots, P_?)$

③ 确定性 $J_c(1,0,\cdots,0) = J_c(0,1,\cdots,0) = \cdots = 0$

④ 扩张性 $J_c(P_1, \cdots, P_c) = J_{c+1}(P_1, \cdots, P_c, 0)$

⑤ 连续性 $P(\omega_i \mid x)$ 的连续函数

Shannon熵: $H = -\sum_{i=1}^{c} P(\omega_i \mid x) \log_2 P(\omega_i \mid x)$

平方熵: $H = 2\left[1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)\right]$

熵可分离性判据: $J_e = \int H(x)p(x)dx$

J_e 大，则重叠性大，可分性不好。

J_e 小，则可分性好。

类别可分性测度——应用举例

图像分割：Otsu灰度图像阈值算法 (Otsu thresholding) —— **最大类间方差法**

给定一幅图像，其归一化直方图表示为： $p_i = n_i / N;\quad p_i \geq 0;\quad \sum_i p_i = 1$

设分割的阈值为k，则背景和目标类的概率为： $\omega_0 = \Pr(C_{bg}) = \sum_{i=0}^{k} p_i = \omega(k)$ $\omega_1 = \Pr(C_{obj}) = \sum_{i=k+1}^{L-1} p_i = 1 - \omega(k)$

背景类的均值为： $\mu_0 = \sum_{i=0}^{k} i \Pr(i \mid C_{bg}) = \sum_{i=0}^{k} i \Pr(i, C_{bg}) / \Pr(C_{bg}) = \sum_{i=0}^{k} i\, p_i / \omega_0 = \mu(k) / \omega(k)$ $\mu(k) = \sum_{i=0}^{k} i p_i$

目标类地均值为： $\mu_1 = \sum_{i=k+1}^{L-1} i \Pr(i \mid C_{obj}) = \sum_{i=k+1}^{L-1} i\, p_i / \omega_1 = \frac{\mu_T - \mu(k)}{1 - \omega(k)}$ $\mu_T = \sum_{i=0}^{L-1} i p_i$

$\omega_0, \omega_1, \mu_0, \mu_1$ 满足： 图像均值： $\omega_0 \mu_0 + \omega_1 \mu_1 = \mu_T$

$\omega_0 + \omega_1 = 1$

类别可分性测度——应用举例

图像分割：Otsu灰度图像阈值算法 (Otsu thresholding) —— **最大类间方差法**

背景类的方差 $\sigma_0^2=\sum_{i=0}^{k}(i-\mu_0)^2\Pr(i\mid C_{bg})=\sum_{i=0}^{k}(i-\mu_0)^2 p_i/\omega_0$

目标类的方差 $\sigma_1^2=\sum_{i=k+1}^{L-1}(i-\mu_1)^2\Pr(i\mid C_{obj})=\sum_{i=k+1}^{L-1}(i-\mu_1)^2 p_i/\omega_1$

类内方差为： $\sigma_w^2=\omega_0\sigma_0^2+\omega_1\sigma_1^2$

类间方差为： $\sigma_B^2=\omega_0(\mu_0-\mu_T)^2+\omega_1(\mu_1-\mu_T)^2=\omega_0\omega_1(\mu_1-\mu_0)^2$

寻找使得类间方差最大的阈值

参考文献：K, Fukunage, Introduction to Statisticul Pattern Recogniition. New York: Academic, 1972, pp. 260-267.

类别可分性测度——应用举例

图像分割：Otsu灰度图像阈值算法 (Otsu thresholding) —— **最大类间方差法**

步骤1：

```
BYTE* ptr = pimg;
for (i=0; i<imsize; i++)          // 统计直方图
        histogram[*ptr++]++;
```

步骤2：

```
float prob[256], miu[256], miuT = 0;
for (i=0; i<256; i++) {
        prob[i]= (float)histogram[i] / imsize;    // 各灰度级的概率
        miuT += miu[i] = i * prob[i];             // 各灰度级的质量矩，
}
```

$$\mu_T = \sum_{i=0}^{L-1} ip_i$$

类别可分性测度——应用举例

图像分割：Otsu灰度图像阈值算法 (Otsu thresholding) —— **最大类间方差法**

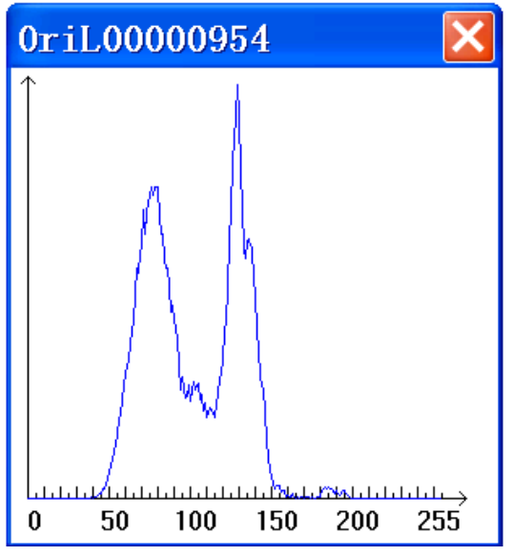
步骤3：寻找最大类间方差

```
BYTE t = 0;          // 阈值t
float miu0 = 0, miu1, miuk = 0, wk = 0, w0, w1, sigma, sigma_max = -1;
for(i=0; i<256; i++) {
    wk += prob[i];    miuk += miu[i];
    w0 = wk;    w1 = 1 - wk;
    miu0 = miuk / w0;
    miu1 = (miuT - miuk) / w1;
    sigma = w0*w1*(miu1-miu0)*(miu1-miu0);
    // 寻找最大sigma值
    if ( sigma >= sigma_max ) {
        t = i;
        sigma_max = sigma;
    }
}
```

$$\mu_0 = \mu(k)/\omega(k)$$

$$\mu_1 = \frac{\mu_T - \mu(k)}{1-\omega(k)}$$

$$\sigma_B^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2$$



- **特征选择与提取基本概念**
  基本概念，特征形成，特征评价准则…
- **类别可分性测度**
  基于距离、概率分布、熵函数的可分判据 …
- **特征选择的最优搜索方法**
  穷举，分支定界，遗传算法 …
- **特征提取方法**
  基于类别可分离性判据的特征提取 …
- **主成分分析**
  PCA基本原理 …

特征选择的最优搜索方法

这里只讨论
特征选择

特征选择：从 D 个特征中选出 $d\,(<D)$个

特征提取：把 D 个特征变为 $d\,(<D)$ 个新特征

要求：从 D 个特征中选出 d 个，应该如何选，使分类性能最佳（J 最大）

从原始特征中挑选出一些最有代表性、分类性能最好的特征进行分类

两个问题： 一是标准， 二是算法

根据实际情况选择某种判据

搜索问题

组合数 $C_D^d = \dfrac{D!}{(D-d)!d!}$

e.g. D=100, d=2, C = 4950

D=100, d=10, C = 1.73103e+13

D=1000, d=2, C = 499500

D=100, d=3, C = 161700

D=100, d=50, C = 1.00891e+29

D=10000, d=2, C = 4.9995e+07

穷举搜索——最优

非穷举搜索——次优

搜索方向：从底向上（特征数从零逐步增加到d ）

从顶向下（特征数从D开始逐步减少到d ）

特征选择的最优搜索方法

穷举法： 计算出各种可能特征组合的某个测度值，加以比较，选择最优特征组。

特点： 可以得到最优特征组；

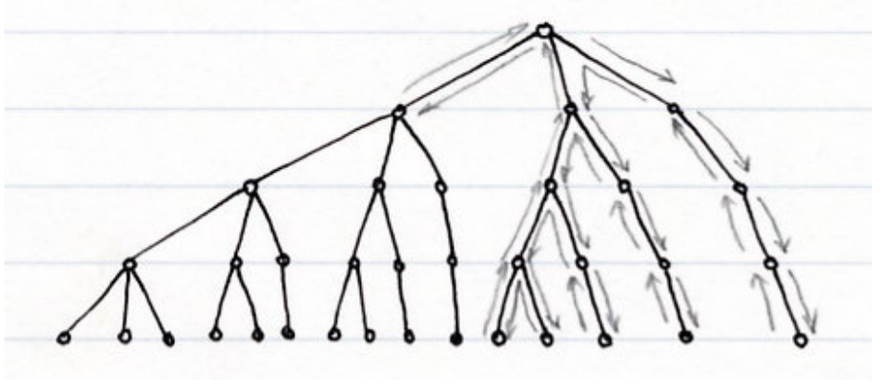
计算量大，难实现。采取搜索技术可降低计算量。

适用于： d 或 D-d 很小（组合数较少）的情况。

分支定界算法： 从顶向下，有回溯

应用条件：利用准则函数的单调性，

基本思想：按照一定的顺序将所有可能的组合排成一棵树，沿树进行搜索，避免一些不必要的计算，使找到最优解的机会最早。



特征选择的最优搜索方法

例：6个特征中选择2个特征

节点上标注的数字为去掉的特征序号

每一级在上一级基础上再去掉一个特征

分支定界算法

方法：

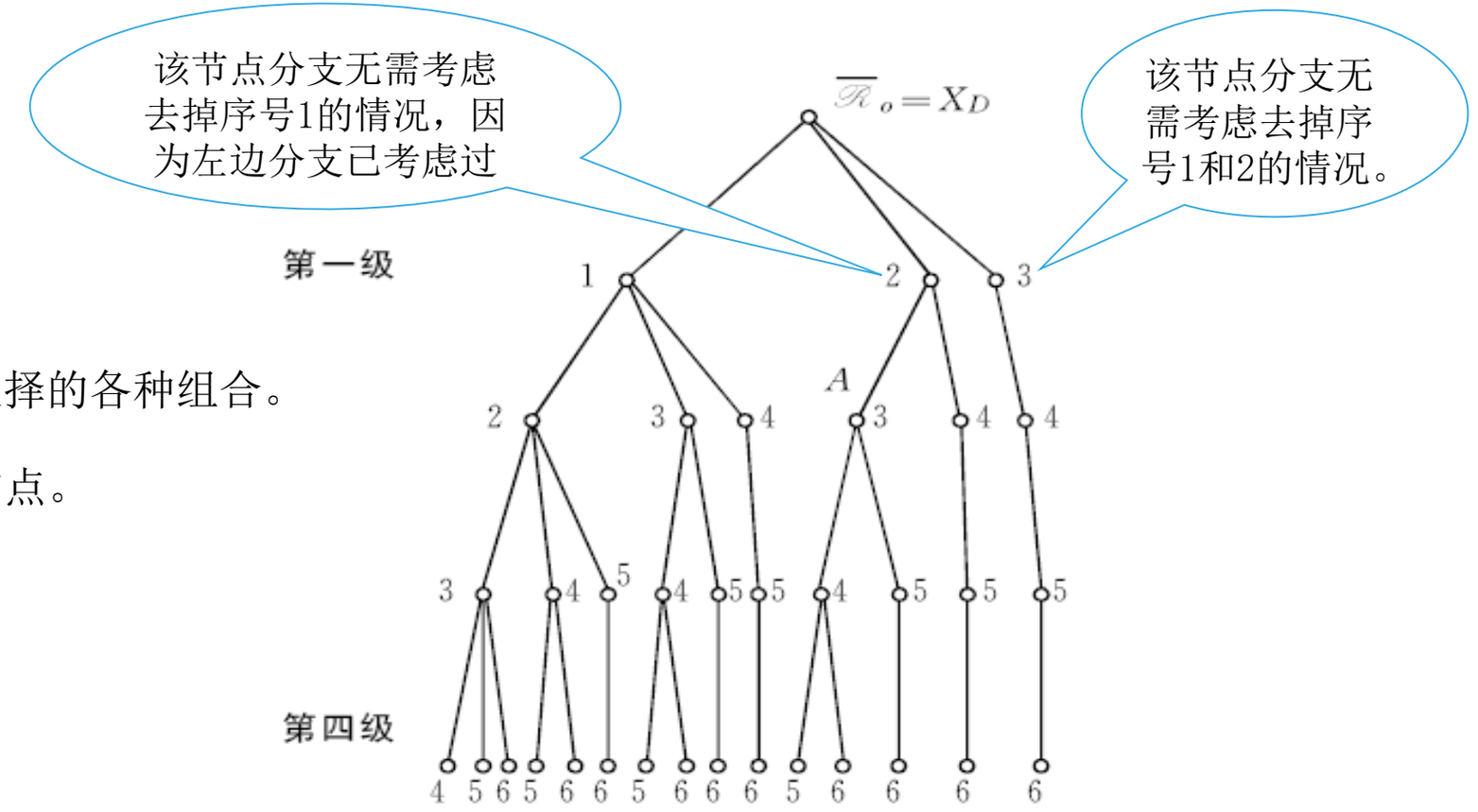
第一步：构建状态图

根结点为第0级，包含全体特征，

每个结点上舍弃一个特征，各叶结点代表选择的各种组合。

避免在整个树中出现相同组合的树枝和叶结点。

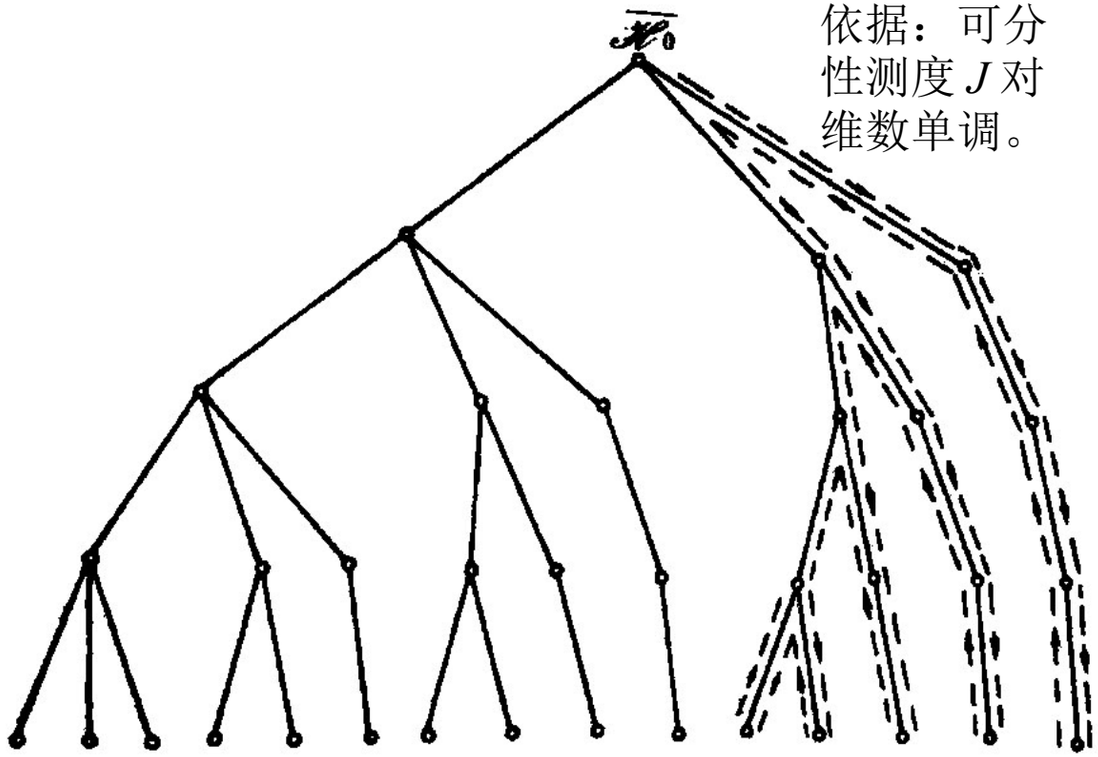
因此搜索树的分支呈现是**左密右疏**的情况。



特征选择的最优搜索方法

例：6个特征中选择2个特征

最优搜索过程



依据：可分性测度 J 对维数单调。

根据类别可分性测度的性质：

$$J_{ij}\left(x_1,x_2,...,x_d\right)\le J_{ij}\left(x_1,x_2,...,x_d,x_{d+1}\right)$$

分支定界算法

方法：

第二步：最优搜索

■ 从右侧开始搜索，初始化B=0，从右到左的顺序搜索；

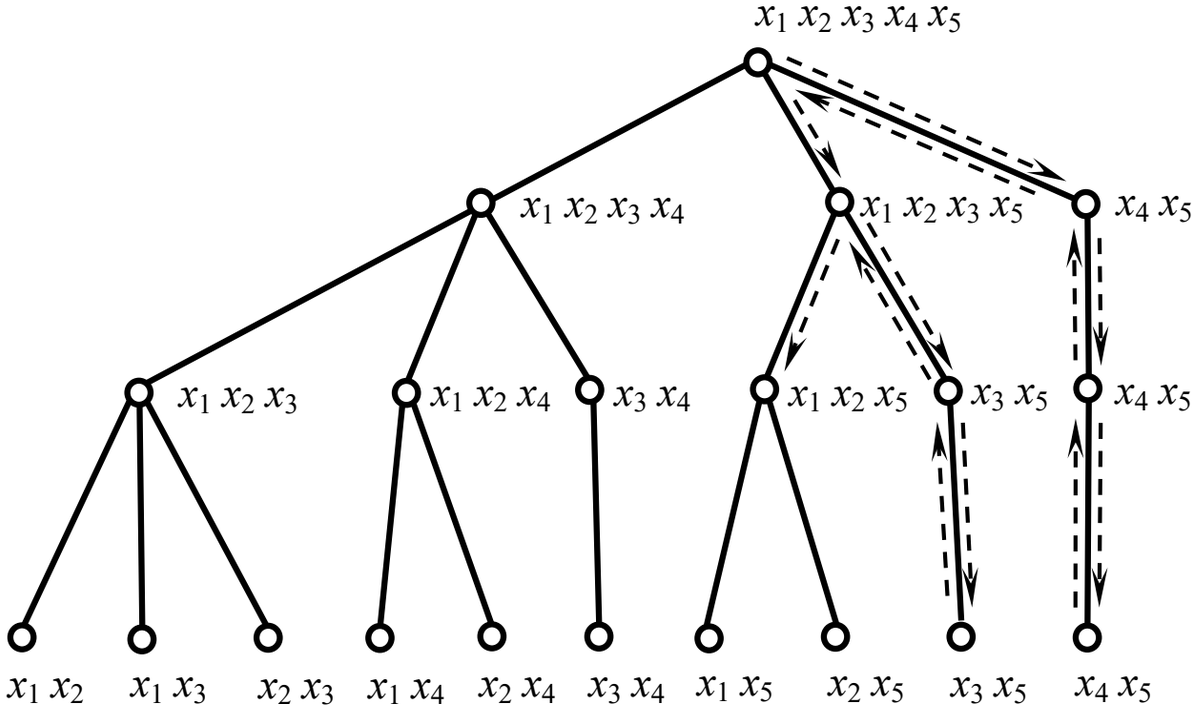
■ 若当前节点没有分支，则向下搜索，直到叶节点，计算可分性依据***J***，若***J>B***，则更新***B***，并将该特征集合作为当前最优选择；向上回溯，按从右到左搜索其他未搜索过的节点。

■ 若当前节点有分支，计算当前节点代表的特征集合的可分性判据 ***J***，若 ***J<B*** 则中止该节点向下搜索，因为其子节点的可分性判据不可能大于 ***J***。否则，按从右到左搜索其子节点。

特征选择的最优搜索方法

分支定界算法

例：从5个特征中选出2个特征作为模式向量。



特征选择的次优搜索方法

虽然不能得到最优解，但可减少计算量。

1）单独最优特征组合：选前d个单独最佳的特征；

计算各特征单独使用时的可分性判据 J 并加以排队，取前d个作为选择结果

2）顺序前进法（Sequential Forward Selection，SFS）

从底向上，每加入一个特征寻优一次，使加入该特征后所得组合最大，

特点：考虑了特征间的相关性，但某特征一经入选，则无法淘汰；

3）广义顺序前进法（Generalized SFS，GSFS）

从底向上，每次增加l个特征。考虑了新增特征中的相关性，

计算量比SFS大，若 $l = d$，（一步加满），则就是穷举法；

4）顺序后退法（Sequential Backward Selection，SBS）

从顶向下，每次减一个特征，与SFS相对，一旦失去，无法换回

特征选择的次优搜索方法

5）广义顺序后退法（Generalized SBS，GSBS）：　从顶向下，每次减 r 个特征

6）增 l 减 r 法（$l-r$法）：SFS和SBS的组合。　自底向上，每次增 $\boldsymbol{l}$ 个再减 $\boldsymbol{r}$ 个特征（$l>r$）

或向顶向下，每次减 $\boldsymbol{r}$ 个再增 $\boldsymbol{l}$ 个特征（$l<r$）

特点：带有局部回溯过程

7）广义的$l-r$ 法（（z_l, z_r）法）　增 l 分成 z_l 步进行，减 r 分成 z_r 步进行。

目的是在适当考虑特征间相关性的同时又能保持适当的计算量。

(Z_l, Z_r)算法		等　效　算　法
$Z_l=(1)$	$Z_r=(0)$	SFS(1,0)算法
$Z_l=(0)$	$Z_r=(1)$	SBS(0,1)算法
$Z_l=(d)$	$Z_r=(0)$	穷举法
$Z_l=(l)$	$Z_r=(0)$	GSPS(l)算法
$Z_l=(0)$	$Z_r=(r)$	GSBS(r)算法
$Z_l=(1,1,\cdots,1)$	$Z_r=(1,1,\cdots,1)$	(l,r)算法

特征选择的随机搜索方法

模拟退火（Simulated Annealing）算法

Tabu搜索（Tabu Search）算法

遗传算法（Genetic Algorithm）

模拟退火（Simulated Annealing）算法

来源于统计力学。材料粒子从高温开始，非常缓慢地降温(退火)，粒子就可在每个温度下达到热平衡。

假设材料在状态 i 的能量为 $E(i)$，那么材料在温度 T 时从状态 i 进入状态 j 遵循如下规律：

如果$E(j) \leq E(i)$，接受该状态被转换。

如果$E(j)>E(i)$，则状态转换以如下概率被接受：

$$e^{\frac{E(i)-E(j)}{KT}}$$

其中K是物理学中的玻尔兹曼常数，T是材料的温度。

特征选择的随机搜索方法

模拟退火（Simulated Annealing）算法

在某一温度下，进行了充分转换后，材料达到热平衡，这时材料处于状态 i 的概率满足玻尔兹曼分布：

$$P_T(x=i)=\frac{e^{-\frac{E(i)}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)}{KT}}}$$

其中 x 是材料当前状态的随机变量，S 为状态空间集合。

显然，所有状态在高温下具有相同概率。 $\lim_{T\to\infty}\frac{e^{-\frac{E(i)}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)}{KT}}}=\frac{1}{|S|}$

其中 $|S|$ 表示集合 S 中状态的数量。

当温度降至很低时，材料会以很大概率进入最小能量状态。 $\lim_{T\to 0}\frac{e^{-\frac{E(i)-E_{\min}}{KT}}}{\sum_{j\in S} e^{-\frac{E(j)-E_{\min}}{KT}}}=\begin{cases}\frac{1}{|S_{\min}|} & i\in S_{\min}\\ 0 & otherwise\end{cases}$

特征选择的随机搜索方法

模拟退火（Simulated Annealing）算法

模拟退火步骤：

Step1: 令i=0, k=0, 给出初始温度T_0和初始特征组合$x(0)$。

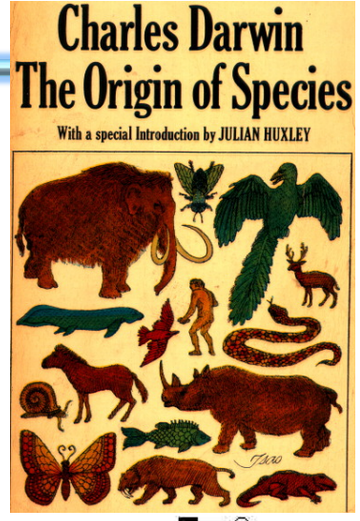
Step2: 在$x(k)$的邻域$N(x(k))$中选择一个状态x'，即新特征组合。计算其可分性判据$J(x')$，并按概率P接受$x(k+1)=x'$。

Step3: 如果在T_i下还未达到平衡，则转到Step2。

Step4: 如果T_i已经足够低，则结束，当时的特征组合即为算法的结果。否则继续。

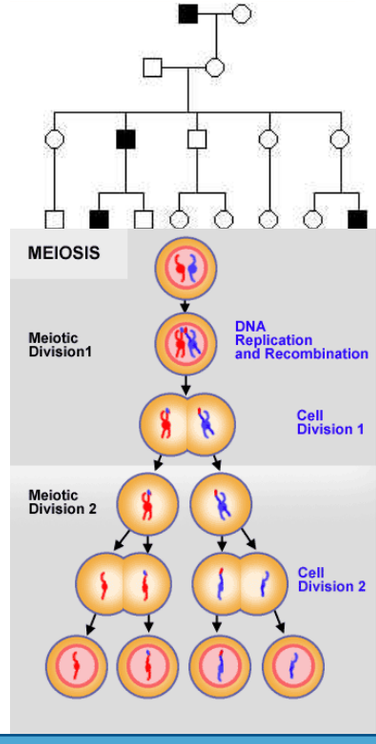
Step5: 根据温度下降方法计算新的温度T_{i+1}。转到Step2。

特征选择的随机搜索方法



遗传算法（Genetic Algorithm）

从生物进化论得到启迪。遗传，变异，自然选择。基于该思想发展了遗传优化算法。



诞生

1、1967年，Holland学生J. D. Bagley在博士论文中首次提出“遗传算法（Genetic Algorithms）”；

2、 1971年，R.B.Hollstien在他的博士论文中首次把遗传算法用于函数优化；

3、 1975年，Holland出版了他的著名专著《自然系统和人工系统的自适应》（Adaptation in Natural and Artificial Systems），这是第一本系统论述遗传算法的专著；

4、 K.A.De Jong完成了他的博士论文《一类遗传自适应系统的行为分析》（An Analysis of the Behavior of a Class of Genetic Adaptive System）

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本术语

个体

- 个体就是模拟生物个体，对问题中对象（一般就是问题的解）的一种称呼。
- 一个个体也就是搜索空间中的一个点。

种群

- 种群(population)就是模拟生物种群, 由若干个体组成的群体.
- 它一般是整个搜索空间的一个很小的子集.

适应度(fitness)

- 借鉴生物个体对环境的适应程度, 对问题中的个体对象所设计的表征其优劣的一种测度。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本术语

适应度函数(fitness function)

- 是问题中全体个体与其适应度之间的一个对应关系。
- 通常为实值函数。
- 该函数就是遗传算法中指导搜索的评价函数。

染色体（chromosome）与基因（gene）

- 染色体是问题中个体的某种字符串形式的编码表示。
- 字符串中的字符称为基因。

例如：

个体		染色体
9	----	1001
（2，5，6）	----	010 101 110

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本算子/基本操作

遗传操作亦称遗传算子(genetic operator)：关于染色体的运算。

遗传算法中有三种遗传操作:

- 选择-复制(selection-reproduction)
- 交叉(crossover，亦称交换、交配或杂交)
- 变异(mutation，亦称突变)

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本算子/基本操作

选择-复制

对于一个规模为 N 的种群 S，按每个染色体 $x_i \in S$ 的选择概率$P(x_i)$所决定的选中机会, 分 N 次从 S 中随机选定 N 个染色体, 并进行复制。

这里的选择概率$P(x_i)$的计算公式为:

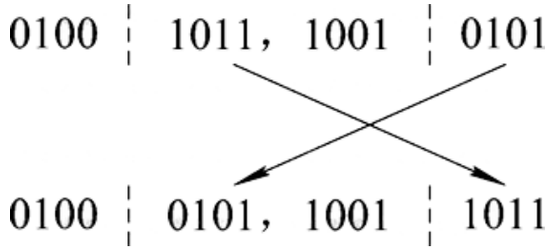
$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)}$$

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本算子/基本操作

交叉: 互换两个染色体某些位上的基因。

例如, 设染色体 *s*1=01001011, *s*2=10010101, 交换其后4位基因, 即



*s*1′=01000101, *s*2′=10011011

可以看做是原染色体*s*1和*s*2的子代染色体。

变异: 改变染色体某个(些)位上的基因。

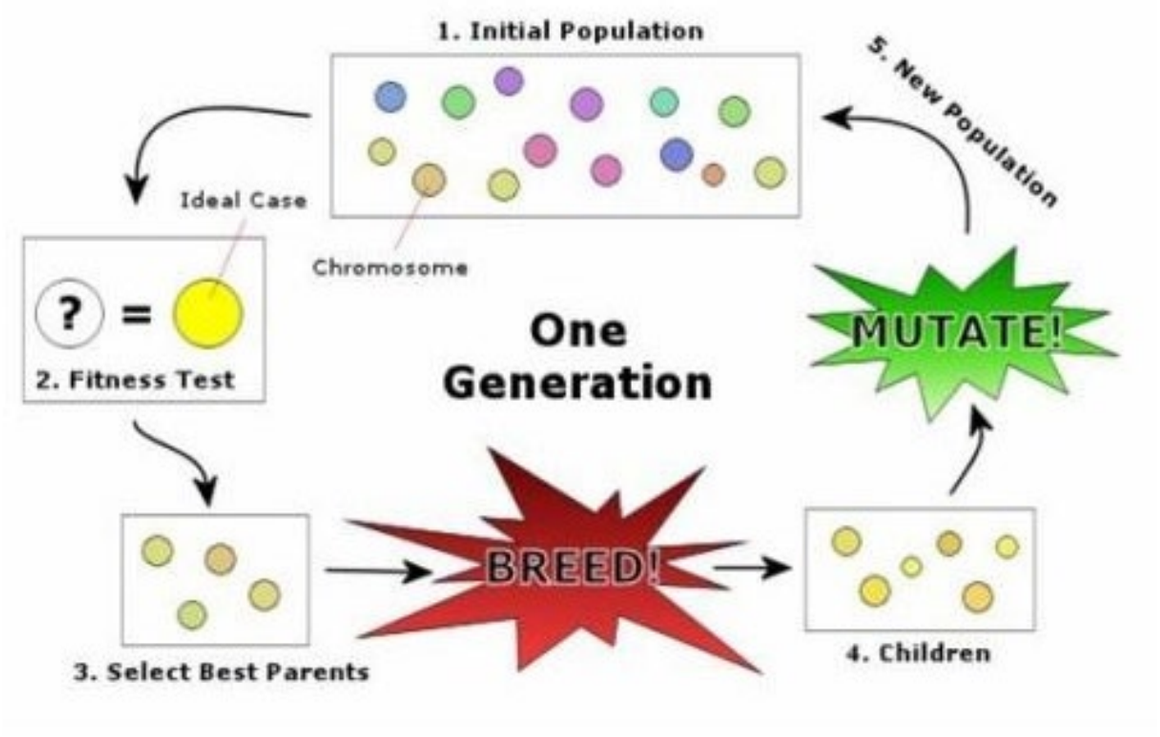
例如, 设染色体 s = 11001101, 将其第三位上的0变为1, 即

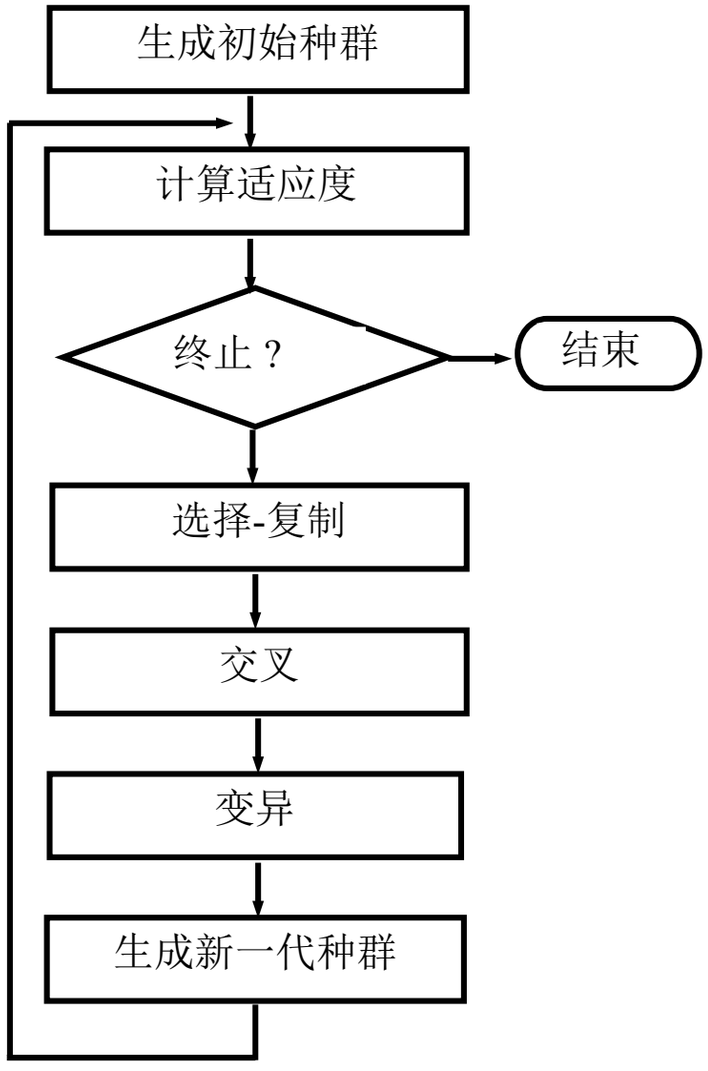
s = 11001101 →11101101= s′。

s′也可以看做是原染色体s的子代染色体。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）基本步骤





特征选择的随机搜索方法

遗传算法（Genetic Algorithm）控制参数

种群规模

最大换代数

交叉率(crossover rate)

- 参与交叉运算的染色体个数占全体染色体总数的比例，记为P_c, 取值范围一般为0.4~0.99。

变异率(mutation rate)

- 指发生变异的基因位数所占全体染色体的基因总位数的比例，记为P_m，取值范围一般为0.0001~0.1。

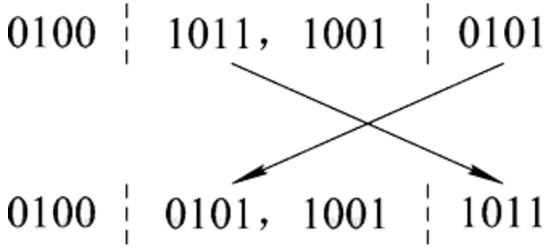
复制的作用：

保留优良个体，但不会产生新个体。

控制进化的方向，而交叉、变异等算子不能控制进化方向。

交叉的作用：

会产生新个体，但子代个体与父代的差异不大。



如上图，只能产生 0100,xxxx,xxxx,xxxx 这个范围内新个体。

与父代的 差异较小，进化速度慢，且易于陷入局部最小。

变异的作用：

会产生突变的新个体，会产生差异较大的新个体

若变异率太大，使得进化变成了随机算法。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）

特征搜索空间 { 连续域 / 离散域

举例：利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

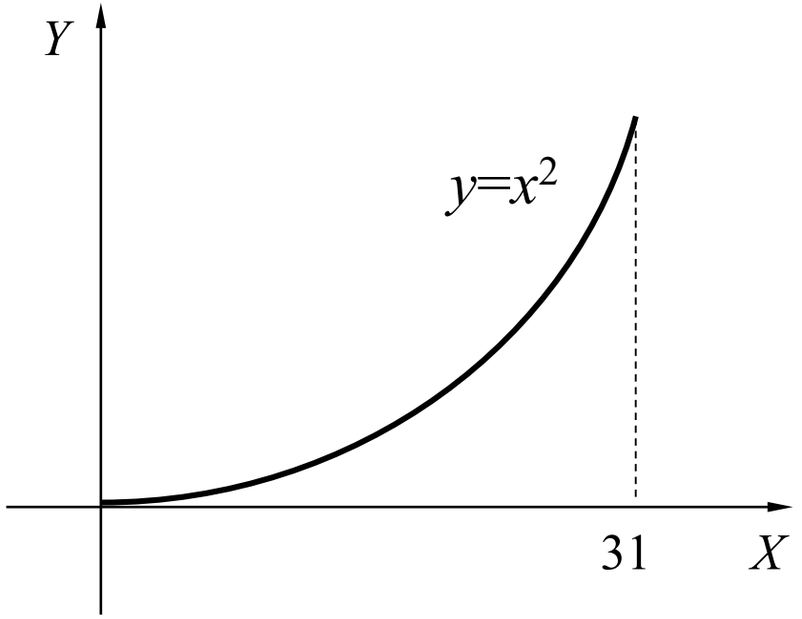
原问题可转化为在区间［0,31］中搜索能使y取最大值的点a的问题。

[0, 31]中的每个点x就是个体。

函数值$f(x)$恰好就可以作为x的适应度。

区间[0, 31]就是一个(解)空间。

问题：如何给出个体x的适当染色体编码？使得该问题可用遗传算法来解决。



特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

(1) 设定种群规模, 编码染色体，产生初始种群。

- 假定将种群规模设定为4；
- 采用5位二进制数编码染色体；
- 随机选取4个个体，组成初始种群S_1:
  - s_1= 13 (01101)
  - s_2= 24 (11000)
  - s_3= 8 (01000)
  - s_4= 19 (10011)

（2）定义适应度函数

- 取适应度函数：$f(x)=x^2$
- 适应度函数选取是具体问题具体分析。

（3）计算当前种群中的各个体的适应度。

- 以初始种群S_1中各个体的适应度$f(s_i)$。
- s_1= 13(01101), s_2= 24(11000)，
- s_3= 8(01000), s_4= 19(10011)
- 有：$f(s_1)=f(13)=13^2=169$

  $f(s_2)=f(24)=24^2=576$

  $f(s_3)=f(8)=8^2=64$

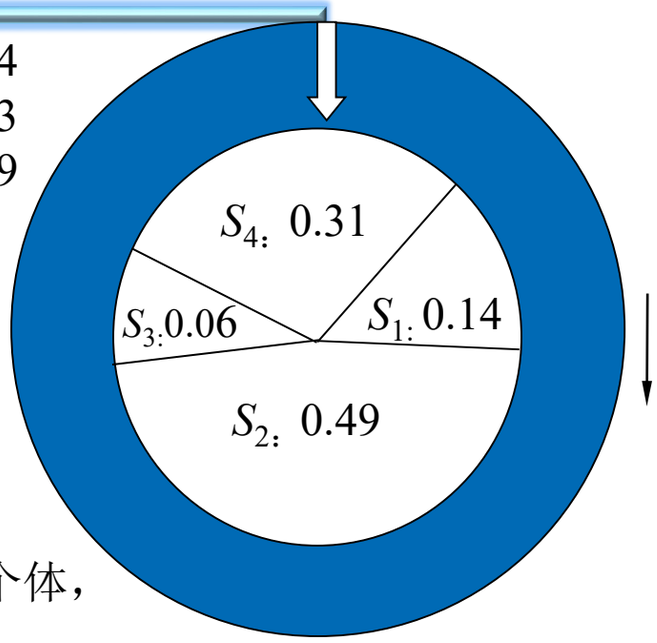
  $f(s_4)=f(19)=19^2=361$

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

S_1: 0 ~ 0.14
S_2: 0.14 ~ 0.63
S_3: 0.63 ~ 0.69
S_4: 0.69 ~ 1



（4）计算当前种群中各个体的选择概率。

- 选择概率的计算公式为 $$P(x_i)=\frac{f(x_i)}{\sum_{j=1}^{N} f(x_j)}$$
- 以初始种群S_1为例：
  - $P(s1)=P(13)=169/1170=0.14$
  - $P(s2)=P(24)=576/1170=0.49$
  - $P(s3)=P(8)=64/1170=0.06$
  - $P(s4)=P(19)=361/1170=0.31$

（5）选择-复制：

- 赌轮选择法选择下一代种群个体，
- 赌轮选择法过程:
- ① 在[0,1]区间内产生一个均匀分布的随机数r。
- ② 若$r \le q_1$,则染色体$x1$被选中。
- ③ 若 $q_{k-1} < r \le q_k (2 \le k \le N)$, 则染色体 x_k 被选中。

  其中的q_i称为染色体x_i (i=1，2，…，n)的**积累概率**，其计算公式为

$$q_i=\sum_{j=1}^{i} P(x_j)$$

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

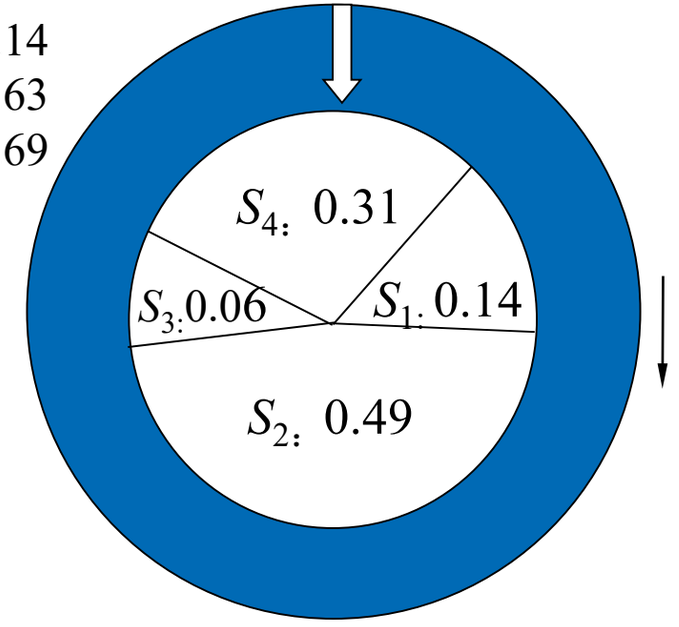
利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

（5）选择-复制：　设从区间[0,1]中产生4个随机数如下:

- $r1 = 0.450126$,　$r2 = 0.110347$
- $r3 = 0.572496$,　$r4 = 0.98503$

染色体	适应度	选择概率	积累概率	选中次数
s_1=01101(13)	169	0. 14	0. 14	1 ($r2$)
s_2=11000(24)	576	0. 49	0. 63	2 ($r1$和$r3$)
s_3=01000(8)	64	0. 06	0. 69	0
s_4=10011(19)	361	0. 31	1. 00	1 ($r4$)

S_1:　0 ~ 0.14
S_2: 0.14 ~ 0.63
S_3: 0.63 ~ 0.69
S_4: 0.69 ~ 1



经复制得下一代新种群的初始个体：

- $s1'$=11000（24），$s2'$=01101（13）
- $s3'$=11000（24），$s4'$=10011（19）

结束了吗？

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

（7）交叉

根据交叉率从新种群选取参与交叉运算的个体。

若交叉率p_c=100%，则全体染色体都参加交叉运算。

随机选择两对染色体参与交叉。

随机选择交叉运算时的交换位数。

举例：

- 设在选择-复制获得新种群后，随机选择$s1'$与$s2'$配对交叉，$s3'$与$s4'$配对交叉。

$s1'$=11000（24），$s2'$=01101（13）

随机确定最后两位交叉，分别交换后两位基因，得新染色体：

- $s1''$=11001（25），$s2''$=01100（12）

同样：$s3'$=11000（24），$s4'$=10011（19）

- 最后两位交叉，得新染色体：
- $s3''$=11011（27），$s4''$=10000（16）

新种群中各个体的染色体：

$s1''$=11001（25），$s2''$=01100（12）

$s3''$=11011（27），$s4''$=10000（16）

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

（8）变异

根据变异率p_m，对新种群中个体的染色体基因进行变异。

新种群中各个体染色体的各个基因，产生随机数，若小于p_m，则变异（反转）。

举例：

- 设变异率p_m =0.001
- 从总体看，共有5×4×0.001=0.02 位基因发生变异。0.02位显然不足1位，所以不会有变异。

（9）下一代种群的产生

第二代种群*S*2：

- s_1=11001（25）, s_2=01100（12）
- s_3=11011（27）, s_4=10000（16）

计算新种群中最大的适应度值

- s_3=11011 $f(s_3)$ = 729
- 大于前一代种群中的最大适应度值

新种群“适者生存”，新种群作为当前种群。为第二代种群*S*2。

返回步骤3，迭代

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

第二代种群　求解步骤3、4

当前种群中染色体S_2	适应度	选择概率	积累概率	估计被选中次数
s_1=11001(25)	625	0.36	0.36	1
s_2=01100(12)	144	0.08	0.44	1
s_3=11011(27)	729	0.41	0.85	1
s_4=10000(16)	256	0.15	1.00	1

步骤5：选择-复制

- 这一轮假设选择-复制操作中，经随机选择，将种群*S*2中的**4个染色体都选中**，则得到新群体：
- *s*1’=11001（25），*s*2’= 01100（12）
- *s*3’=11011（27），*s*4’= 10000（16）

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

第二代种群 求解步骤7、8

$s1'$=11001（25），$s2'$= 01100（12），$s3'$=11011（27），$s4'$= 10000（16）

交叉运算，

- 经随机确定，让$s1'$与$s2'$，$s3'$与$s4'$分别交换后三位基因，得：
- $s1''$=11100（28）, $s2''$= 01001（9）
- $s3''$=11000（24）, $s4''$= 10011（19）

变异运算

- 随机确定，没有变异发生。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

第三代种群

得到新一代种群：

- $s1$=11100（28）, $s2$=01001（9）
- $s3$=11000（24）, $s4$=10011（19）

计算新种群中最大的适应度值

- s_1=11100 $f(s_1)$ = 784
- 大于前一代种群中的最大适应度值

新种群“适者生存”，新种群作为当前种群，为第三代种群$S3$ 。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

第三代种群

求解步骤3、4

染色体	适应度	选择概率	积累概率	估计的选中次数
s_1=11100 (28)	784	0.44	0.44	2
s_2=01001 (9)	81	0.04	0.48	0
s_3=11000 (24)	576	0.32	0.80	1
s_4=10011 (19)	361	0.20	1.00	1

设这一轮的选择-复制结果为：

- $s1'$=11100（28），$s2'$=11100（28）
- $s3'$=11000（24），$s4'$=10011（19）

做交叉运算，让$s1'$与$s4'$，$s2'$与$s3'$分别交换后两位基因，得

- $s1''$=11111（31），$s2''$=11100（28）
- $s3''$=11000（24），$s4''$=10000（16）

这一轮仍然不会发生变异。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）连续域举例

利用遗传算法求解区间[0,31]上的二次函数 $y=x^2$ 的最大值。

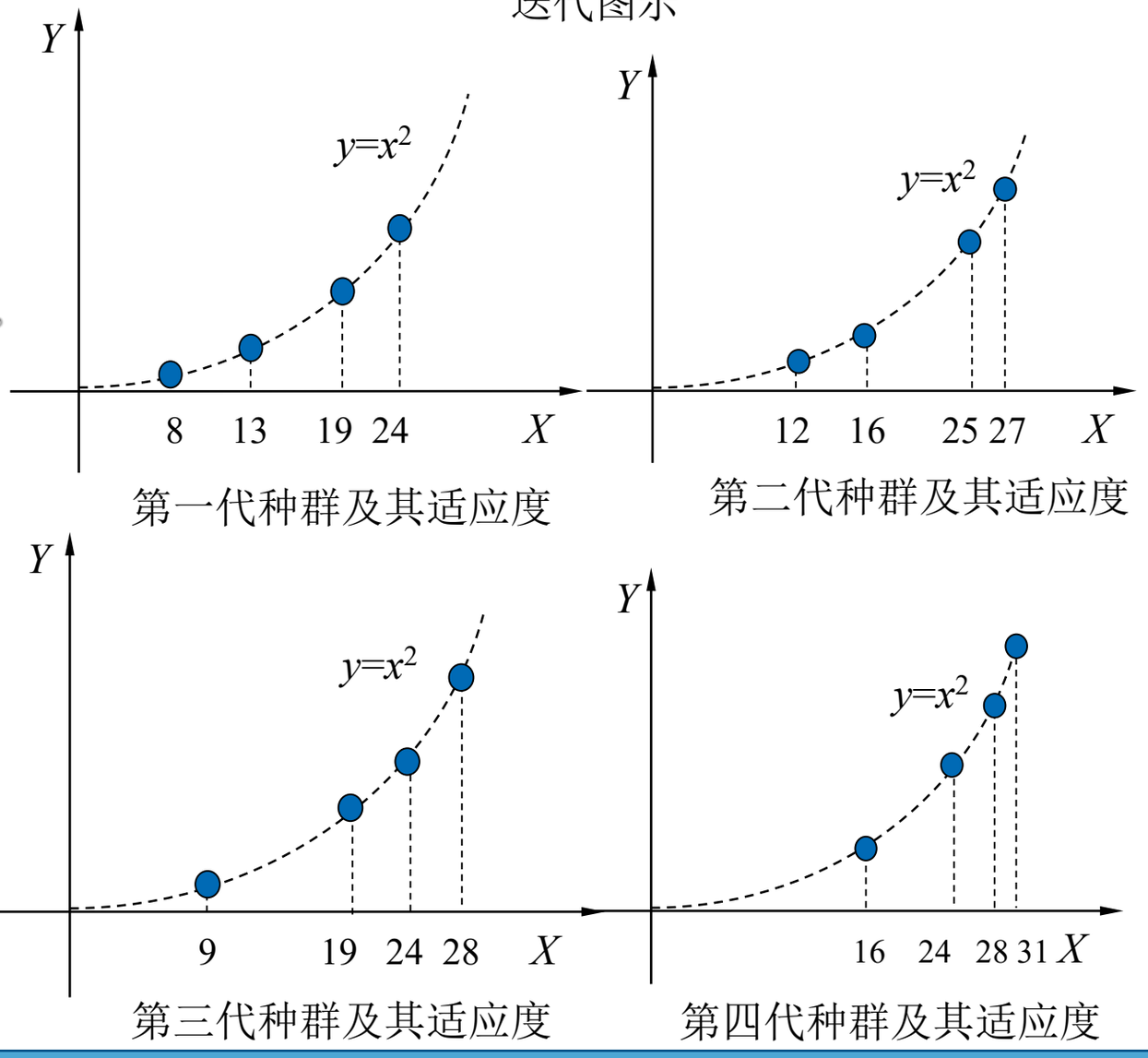
第四代种群

- $s1$=11111（31），$s2$=11100（28）
- $s3$=11000（24），$s4$=10000（16）

在这一代种群中已经出现了适应度最高的染色体$s1$=11111。

之后遗传进化所得的各代种群，最大个体适应度值均小于等于$s1$=11111。于是，遗传操作终止，将染色体“11111”作为最终结果输出。

迭代图示



特征选择的随机搜索方法

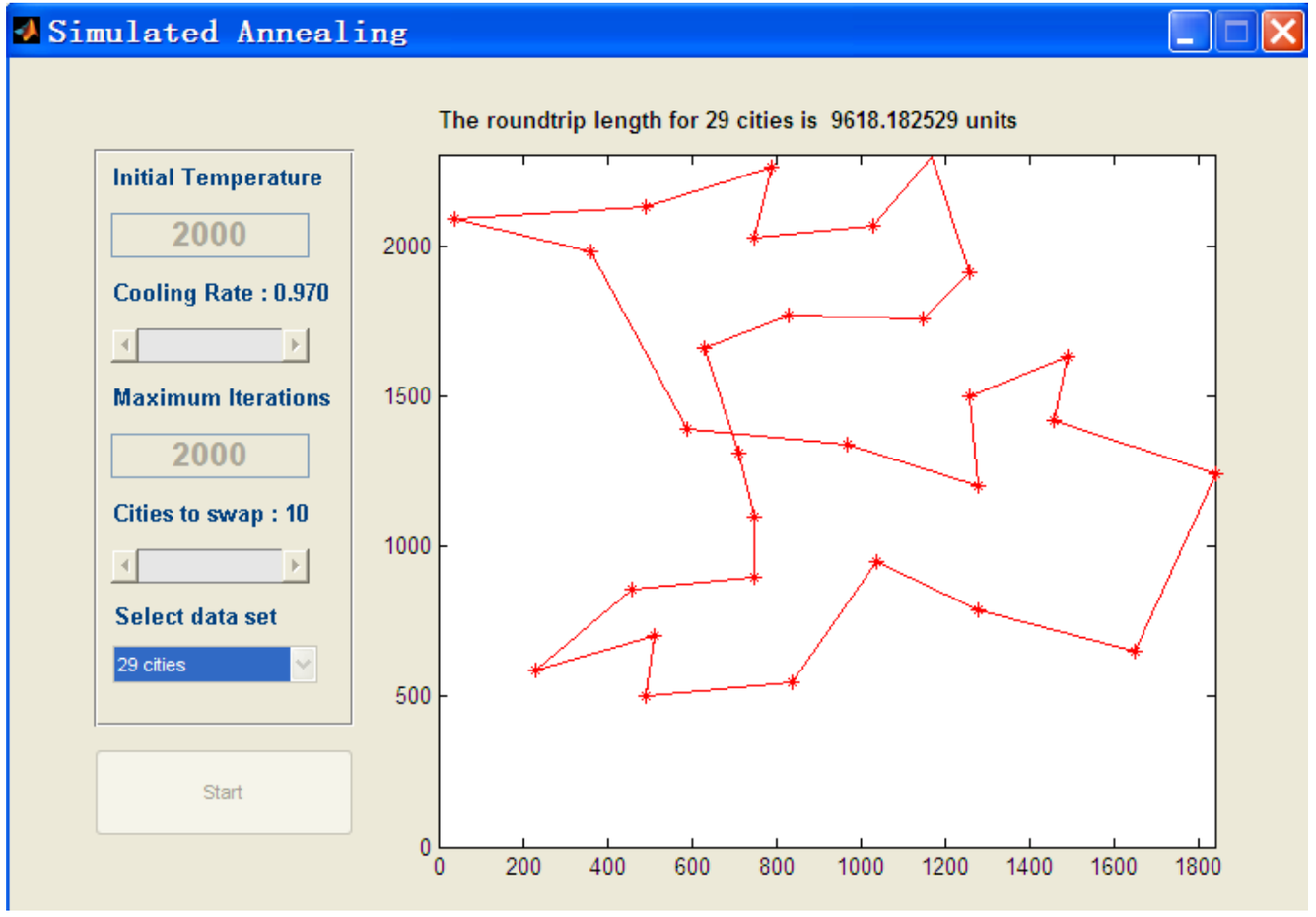
遗传算法（Genetic Algorithm）离散域

特征搜索空间 { 连续域 / **离散域** }

用遗传算法求解TSP

Traveling Salesman Problem

假设有一个旅行商人要拜访n个城市，他必须选择所要走的路径，路径的限制是每个城市只能拜访一次，而且最后要回到原来出发的城市。路径的选择目标是要求得的路径路程为所有路径之中的最小值。



特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

如何构建染色体？

任一可能解：一个合法的城市序列，即n个城市的一个排列。

可直接在解空间（所有合法的城市序列）中搜索最佳解。这正适合用遗传算法求解。

一个合法的城市序列$s=(c_1, c_2, \ldots, c_n, c_{n+1})$ (c_{n+1}就是c_1)作为一个个体。

定义适应度函数

城市序列中相邻两城之间的距离之和的倒数可作为相应个体s的适应度。

即适应度函数就是：$$f(s)=\frac{1}{\sum_{i=1}^{n} d(c_i, c_{i+1})}$$

为什么这么定义？

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

如何编码和交叉？

虽然选定一个合法的城市序列（$c_1, c_2, \ldots, c_n, c_{n+1}$）($c_{n+1}$就是$c_1$)作为一个个体。

还需对个体进行编码。

如果编码不当，就会在实施交叉或变异操作时出现非法城市序列即无效解。

例如，对于5个城市的TSP，用符号A、B、C、D、E代表相应的城市，用这5个符号的序列表示可能解即染色体。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

如何编码和交叉？

设$s1$=(A, C, B, E, D, A)，$s2$=(A, E, D, C, B, A)进行遗传操作:

- 实施常规的交叉操作，如交换后三位，得
  - $s1'$=（A,C,B,C,B,A），$s2'$=（A,E,D,E,D,A）
- 实施常规的变异操作，将染色体$s1$第二位的C变为E，得
  - $s1''$=（A, E, B, E, D, A）
- 可以看出，上面得到的$s1'$，$s2'$和$s1''$都是非法的城市序列。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

解决方法：

对TSP必须设计合适的染色体和相应的遗传运算。

针对TSP已提出了许多编码方法和相应的特殊化的交叉、变异操作，以巧妙地用遗传算法解决TSP。

如顺序编码或整数编码、随机键编码.

如部分映射交叉、顺序交叉、循环交叉、位置交叉。

如反转变异、移位变异、互换变异等等。

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

编码方式：

二进制编码

$$(000\ 001\ 010\ 011\ 100\ 101)$$

- 对TSP而言没有意义，因为城市号就是基本单元。

路径表示

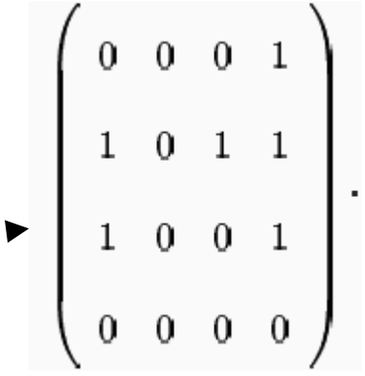
- 最常用的编码方式

(6 3 5 4 1 2)

- 有许多特殊的遗传操作算子

Matrix representation

(2 3 1 4)



特征选择的随机搜索方法

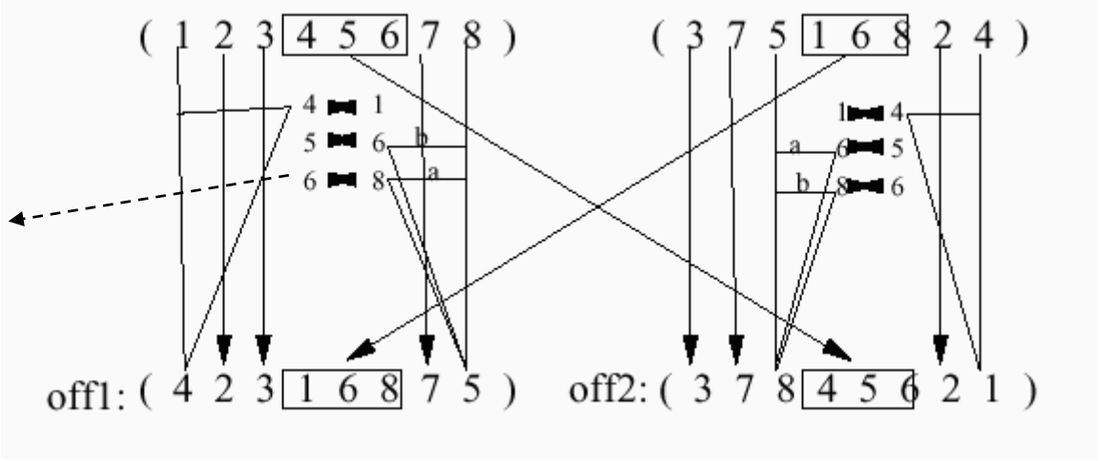
遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP　　交叉方式——部分映射交叉PMX(Partial-Mapped Crossover)：

2 parents　(1 2 3 4 5 6 7 8)　(3 7 5 1 6 8 2 4)

1. 随机取两点,例如4、6
2. 两点内部分先行交换
3. 其余依次填入, 但重复者必须替换

置换规则



1） 第四位交换，4和1交换

原来　1 2 3 4 5 6 7 8　　3 7 5 1 6 8 2 4

左边：4 2 3 1 5 6 7 8　右边：3 7 5 4 6 8 2 1

2） 第五位交换，5和6交换

原来　4 2 3 1 5 6 7 8　右边：3 7 5 4 6 8 2 1

左边：4 2 3 1 6 5 7 8　右边：3 7 6 4 5 8 2 1

3） 第六位交换，6和8交换

原来：4 2 3 1 6 5 7 8　右边: 3 7 6 4 5 8 2 1

左边：4 2 3 1 6 8 7 5　右边: 3 7 8 4 5 6 2 1

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

用遗传算法求解TSP

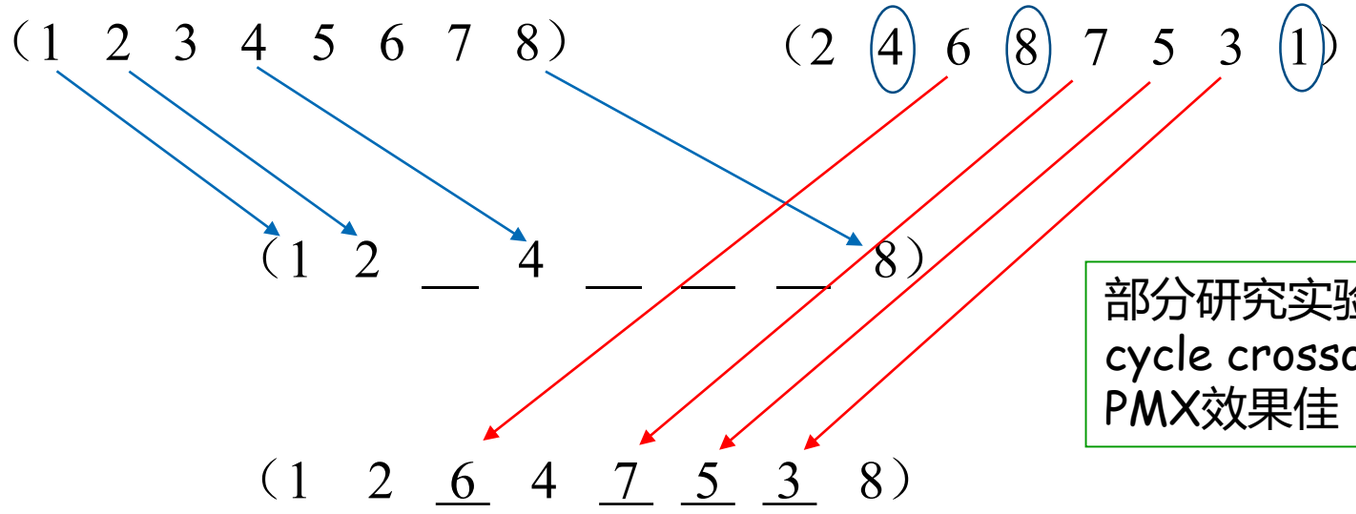
交叉方式

——循环交叉 Cycle Crossover

2 parents 產生一個offspring

(1 2 3 4 5 6 7 8)　(2 4 6 8 7 5 3 1)

1. 从第一个位置开始填起，2 选 1 ，下例为选到左边parent的情形
2. 填好第一个位置之後后，看另一个parent同样number(此处為 1)的位置何在(位置8)
3. 再將原parent同位置的number填入offspring，其餘依此類推
4. 左邊parent沒有number可供填入，則找右邊的parent，從offspring最前面的空位填起



部分研究实验显示cycle crossover较PMX效果佳

特征选择的随机搜索方法

遗传算法（Genetic Algorithm）离散域

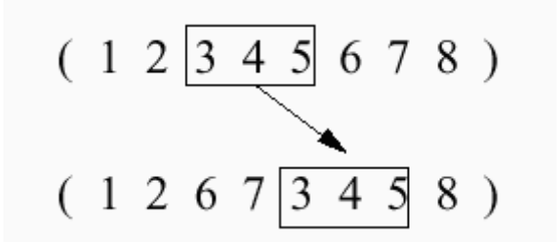
用遗传算法求解TSP

变异方式——移位变异

Displacement Mutation(DM)

the parent (1 2 3 4 5 6 7 8)

1. 随机选出一段sub-tour, 例如选到(3 4 5)
2. 先将其余cities重新排好,如(1 2 6 7 8)
3. 将(3 4 5)随机插入任一個位置

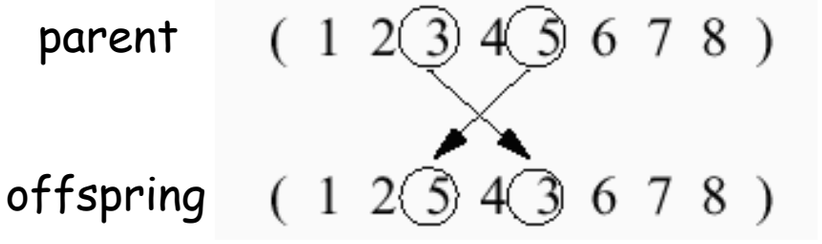


变异方式——互换变异

Exchange Mutation(EM) 或 swap mutation

the parent (1 2 3 4 5 6 7 8)

1. 任意选择兩個位置, 將之交换



Any Questions?