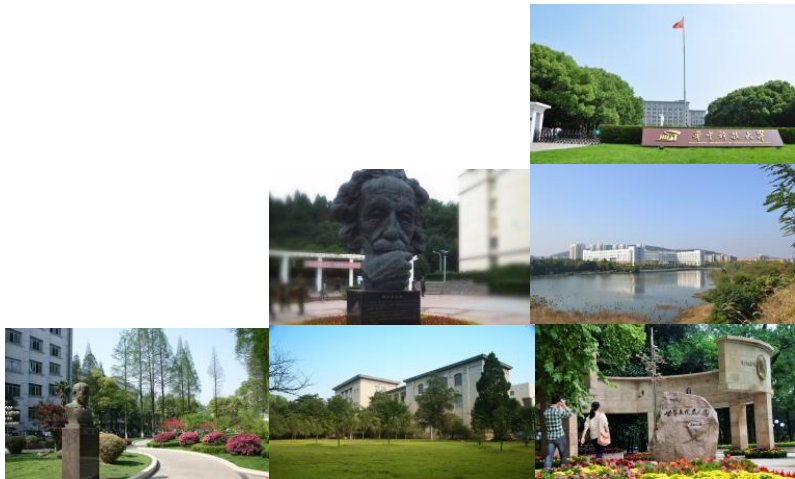
人工智能与自动化学院

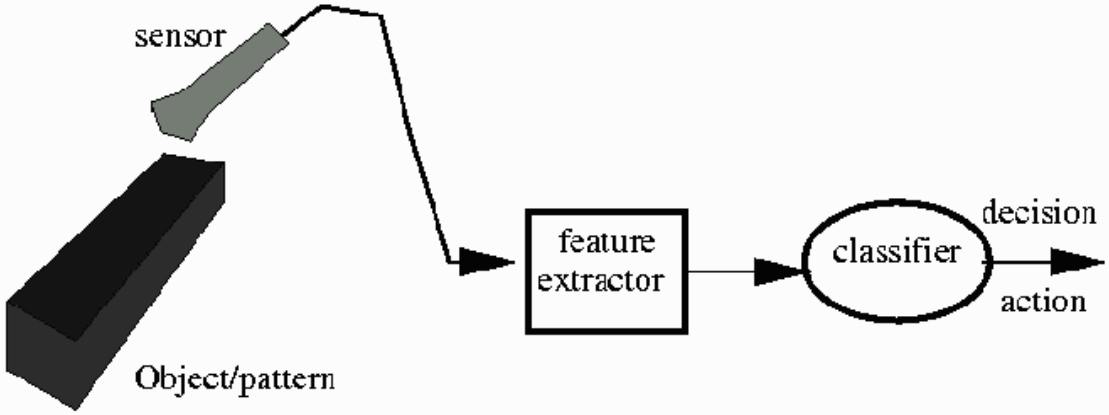
模 式 识 别

特征选择



第五讲　特征选择（*Feature Selection*）

第五讲　特征选择（*Feature Selection*）



样本：即一组特征组成的向量

特征：对事物的观测或其某种运算，用于分类

（模式识别系统的性能非常依赖于特征）

Featured Examples

Image Retrieval Using Customized Bag of Features

Create a Content Based Image Retrieval (CBIR) system using a customized bag-of-features workflow.

全局特征：颜色、纹理、形状

局部特征：SURF、HOG、LBP

引言Introduction

- 特征获取（*Feature acquisition*）

  直接观测到的或经过初步运算的特征——原始特征

- 特征选择（*Feature Selection*）

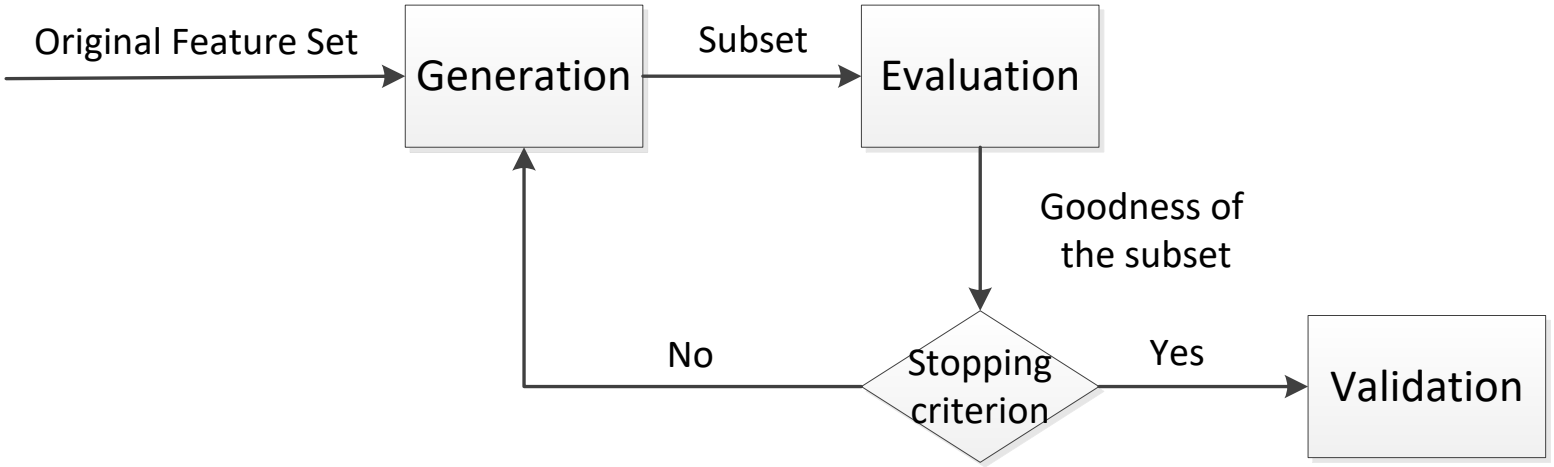
  从 m 个特征中选择 m_1 个，人为选择、算法选择，$m_1 < m$

- 特征提取（*Feature Extraction*）

  特征变换/特征压缩：将 m 个特征变为 m_2 个新特征。

引言Introduction

特征选择的过程



① Generation Procedure：产生过程是搜索特征子集的过程，负责为评价函数提供待评估特征子集。

② Evaluation Function：评价函数是评价一个特征子集好坏程度的一个准则。

③ Stopping Criterion：停止准则是与评价函数相关的，一般是一个阈值，当评价函数值达到这个阈值后就可停止搜索。

④ Validation Procedure：在验证数据集上验证选出来的特征子集的有效性。

5.1特征的评价准则(*Evaluation Criterion*)

评价准则(evaluation criterion)：数学上定义的用以衡量特征对分类的效果的准则。

- 错误率

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

5.1特征的评价准则(*Evaluation Criterion*)

J_{ij} 两类的可分性判据(Two-Class Separability Measures)要求

- 与错误率有单调关系(Monotonic Relationship)。J_{ij}大 $\Rightarrow$ P_e小
- 判据具有“距离”(Distance)的某些特性。$J_{ij} \geq 0$, $J_{ii} = 0$，$J_{ij} = J_{ji}$
- 对独立的特征有可加性(Additivity)。 $J_{ij}(x_1, x_2, \cdots x_d) = \sum_{k=1}^{d} J_{ij}(x_k)$

  式中， $x_1, x_2, \cdots, x_d$ 是对不同种类特征的测量值， $J_{ij}(\cdot)$

  表示使用括号中特征(组)时第i 类与第j 类可分性判据函数。

- 增加特征时判据不减小。 $J_{ij}(x_1, x_2, \cdots, x_d) \leq J_{ij}(x_1, x_2, \cdots, x_d, x_{d+1})$

类均值向量 *(Pooled mean vector)*

某一类的样本集的中心(重心)

$$\vec{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \vec{x}_k^{(i)}$$

总均值向量 *(Total mean vector)*

全体样本的中心(重心)

$$\vec{m} = \sum_{i=1}^{c} P_i \vec{m}_i = \sum_{i=1}^{c} \frac{n_i}{N}\vec{m}_i = \frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{n_i}\vec{x}_k^{(i)} = \frac{1}{N}\sum_{l=1}^{N}\vec{x}_l$$

类间离散度矩阵 S_b 的估计： *(Between-class scatter matrix)*

$$\tilde{S}_b = \sum_{i=1}^{c} P_i(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T$$

类内离散度矩阵 S_w 的估计： *(Within-class scatter matrix)*

$$\tilde{S}_w = \sum_{i=1}^{c} P_i \frac{1}{n_i}\sum_{k=1}^{n_i}\left(\vec{x}_k^{(i)} - \vec{m}_i\right)\left(\vec{x}_k^{(i)} - \vec{m}_i\right)^T = \sum_{i=1}^{c} P_i \Sigma_i$$

Σ_i *(Covariance matrix)*：类协方差阵

可以用 S_W、S_B 构造不同的可分性判据：

各类之间的平均平方距离可以作为判据 $J_1 = tr\left(S_W + S_B\right)$

类内 类间

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

$$J_3 = \ln \frac{|S_b|}{|S_w|}$$ ← 类间 ← 类内

$$J_4 = \frac{\mathrm{tr}S_b}{\mathrm{tr}S_w}$$ ← 类间 ← 类内

类间离散度尽量大

类内离散度尽量小

$$J_5 = \frac{|S_b - S_w|}{|S_w|}$$ ← 类间-类内 ← 类内

$$J_F = \frac{\left(\tilde{m}_1 - \tilde{m}_2\right)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$ Fisher准则函数

特点：直观，易于实现（用样本计算），较常用。
不能确切表明各类分布重叠情况，与错误率无直接联系。
当各类协方差相差不大时，用这些判据较好。

5.1.2 基于概率分布(Probability Distribution)的可分性判据

考查两类分布密度*(Distribution Density)*之间的交叠程度

定义：两个密度函数之间的距离

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

所谓重叠程度是指两个概密函数相似的程度

它必须满足以下条件：

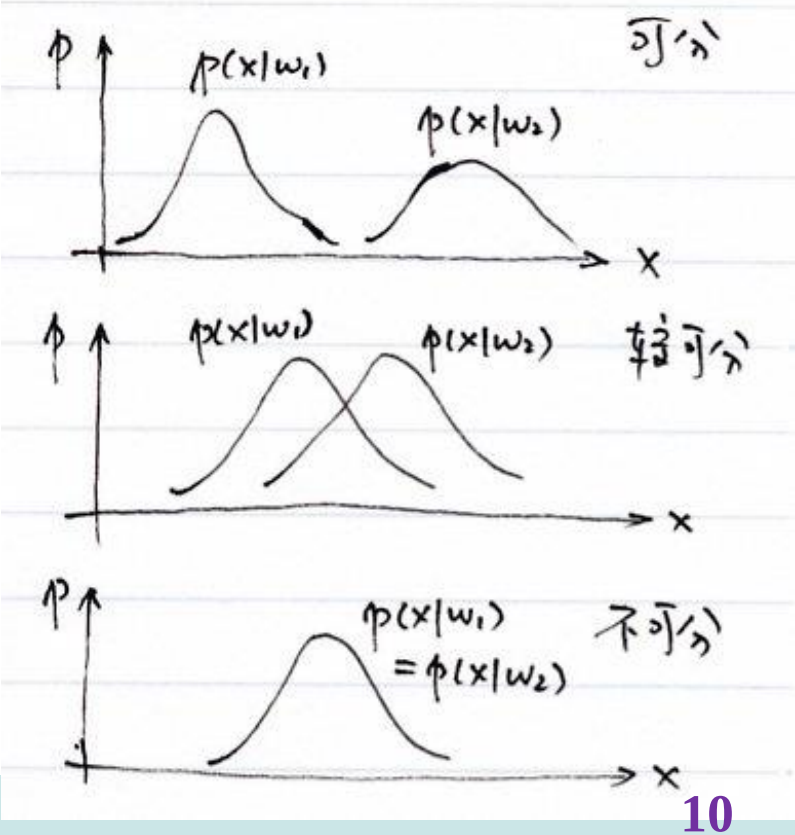
① $J_p \geq 0$

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



5.1.2 基于概率分布(Probability Distribution)的可分性判据

具体定义有多种：

- **Bhattacharyya距离** $J_B = -\ln \int [p(x|\omega_1)p(x|\omega_2)]^{\frac{1}{2}} dx$

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

- **Chernoff界** $J_c = -\ln \int p^s(x|\omega_1)p^{1-s}(x|\omega_2)dx$ $S \in [0, 1]$

当 $s = 0.5$ 时，$J_c = J_B$

与错误率的关系：$P_e \le [P(\omega_1)P(\omega_2)]^{\frac{1}{2}} \exp\{-J_B\}$

- **散度(Divergence)** $J_D = \int_x [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx$

（从负对数似然比发展来的）与错误率有单调关系

正态分布情况下：$J_D = \frac{1}{2} tr[\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I] + \frac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2)$

若 $\Sigma_1 = \Sigma_2 = \Sigma$ （两类协方差阵相等），则

$$8J_B = J_D = (\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) = J_M$$ **Mahalanobis距离平方**

5.1.2 基于概率分布(Probability Distribution)的可分性判据

使用方法

- 由J_B、J_C、J_D的定义式结构以及它们与误分概率的关系可以知道，所选取的特征矢量应使所对应的J_B、J_C 、J_D尽量大，这样可分性就较好。
- 对于c类问题，可采用平均B-判据、C-判据、D-判据

$$\bar{J}_B = \sum_{i=1}^{c} \sum_{j=i+1}^{c} P(\omega_i)P(\omega_j)J_B(\omega_i,\omega_j)$$

$$\bar{J}_C = \sum_{i=1}^{c} \sum_{j=i+1}^{c} P(\omega_i)P(\omega_j)J_C(\omega_i,\omega_j)$$

$$\bar{J}_D = \sum_{i=1}^{c} \sum_{j=i+1}^{c} P(\omega_i)P(\omega_j)J_D(\omega_i,\omega_j)$$

大盖小问题

在特征空间中，若有某两类间的J_B、J_C或J_D很大，可使平均判据变大，这样就掩盖了某些类对的判据值较小的情况存在，从而可能降低总的分类正确率，即所谓的大盖小问题。为改善这种情况，可对每个类对的判据采用变换的方法，使对小的判据较敏感。例如，对J_D，可采用变换

$$\tilde{J}_D(\omega_i,\omega_j)=1-\exp\left[-J_D(\omega_i,\omega_j)/8\right]$$

这样，当ω_i和ω_j两类模式相距很远时，$J_D(\omega_i,\omega_j)$变得很大，但也只能接近于1。但对于散度$J_D(\omega_i,\omega_j)$小的情况，$\tilde{J}_D(\omega_i,\omega_j)$又变得较敏感。于是，总的平均(变换)判据为$\tilde{J}_D(\omega_i,\omega_i)$。

$$\tilde{J}_D=\sum_{i=1}^{c}\sum_{j=i+1}^{c}P(\omega_i)P(\omega_j)\tilde{J}_D(\omega_i,\omega_j)$$

5.1.3 基于熵(Entropy)的可分性判据

熵(Entropy)：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

5.1.3 基于熵(Entropy)的可分性判据

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 $J_c\left(\frac{1}{c}, \cdots, \frac{1}{c}\right) = 1$ **等概率场熵最大，识别最困难。**

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

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

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

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

⑤连续性 (*Continuity*) $P(\omega_i \mid x)$的连续函数

⑥分枝性（综合性）一分为二，则熵增加；二合为一则熵减小

5.1.3 基于熵(Entropy)的可分性判据

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。

5.1.4 用统计检验(statistical test)作为可分性判据

----- 选择在两类间有显著差异的特征

t-test（正态分布假设下）

X	Y
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

sample 1: $X_1,\cdots,X_n \quad X \sim N(\mu_X,\sigma^2)$

sample 2: $Y_1,\cdots,Y_m \quad Y \sim N(\mu_Y,\sigma^2)$

pooled sample variance $s_P^2 = \dfrac{(n-1)S_X^2+(m-1)S_Y^2}{m+n-2}$

the null hypothesis

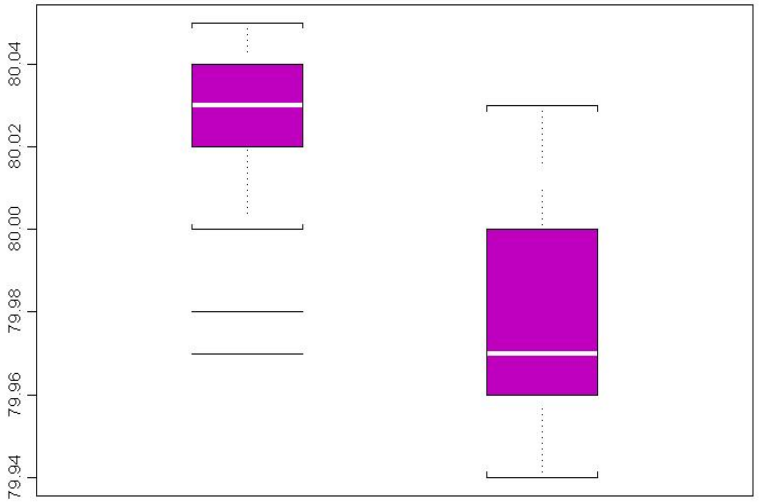
$$H_0: \quad \mu_X = \mu_Y$$

alternative hypotheses

two-sided $H_1: \quad \mu_X \neq \mu_Y$

one-sided $H_2: \quad \mu_X > \mu_Y$

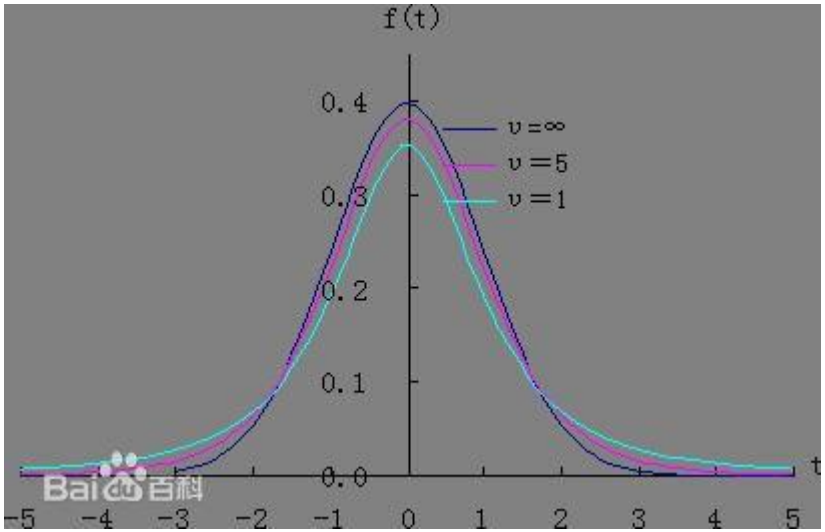
$H_3: \quad \mu_X < \mu_Y$



t-statistic $t = \dfrac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}} = \dfrac{\bar{X}-\bar{Y}}{s_P\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$

5.1.4 用统计检验(statistical test)作为可分性判据

- ✓ t-分布（t-distribution），用于根据小样本来估计呈正态分布且方差未知的总体的均值。如果总体方差已知（例如在样本数量足够多时），则应该用正态分布来估计总体均值。
- ✓ t分布是一簇曲线，其形态与n（确切地说与自由度df）大小有关。与标准正态分布曲线相比，自由度df越小，t分布曲线愈平坦，曲线中间愈低，曲线双侧尾部翘得愈高；自由度df愈大，t分布曲线愈接近正态分布曲线，当自由度df=∞时，t分布曲线为标准正态分布曲线。



$$t=\frac{\bar{X}-\bar{Y}}{s_{\bar{X}-\bar{Y}}}=\frac{\bar{X}-\bar{Y}}{s_P\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad s_P^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{m+n-2}$$

独立样本t检验统计量为

$$t=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n-1)S_X^{\ 2}+(m-1)S_Y^{\ 2}}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

例.采用t检验比较男、女儿童身高是否存在差别。为了进行独立样本t检验，设有自变量（性别：男女）与因变量（测量值）。假设H0：男平均身高=女平均身高，H1：男身高！=女平均身高

① 选用双侧检验，选用alpha=0.05的统计显著水平。

② 数据的排列

被试	性别	身高
对象1	男性	111
对象2	男性	110
对象3	男性	109
对象4	女性	102
对象5	女性	104
	男性身高均数 = 110 女性身高均数 = 103	

$$t=\frac{\bar{X}-\bar{Y}}{s_{X-Y}}=\frac{\bar{X}-\bar{Y}}{s_P\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad s_P^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{m+n-2}$$

$$t=\frac{110-103}{\sqrt{\frac{(3-1)0.667+(2-1)1}{3+2-2}\left(\frac{1}{3}+\frac{1}{2}\right)}}=8.695 \qquad \text{自由度}=n+m-2=3$$

H0的拒绝域：$\left|t\right|\geq t_{\frac{\alpha}{2}}(n-1)$

t 分布表

n	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032

③ 计算独立样本， t-test。选择双侧检验，以及统计显著性水平alpha=0.05。

④ 从输出结果查看t检验的p值，是否达到显著水平。是，接受H1。男平均身高与女平均身高不同。否，接受H0，尚无证据支持男女身高差异。

5.1.4 用统计检验(statistical test)作为可分性判据

t-检验：参数化检验方法，对数据分布有一定假设

非参数检验：不对数据分布作特殊假设

如：**Wilcoxon秩和检验（Rank-sum test）基本做法：**

- 把两类样本混合在一起，按所考查的特征从小到大排序
- 如果一类样本排序序号之和（秩和）显著地比另一类样本小（或大），则两类样本在所考查的特征上有显著差异

5.2 特征选择的最优搜索方法(*Optimal search method*)

问题

从D维特征中选取 d维（$d < D$），

使分类性能最佳（ J 最大）。

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=3, *C* =161700

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

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

D=1000, d=2, *C* = 499500

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 $X_0 = \phi$

从顶向下 $X_0 = X$

5.2 特征选择的最优搜索方法(Optimal search method)

穷举算法(Exhaustive Search/Brute-force search)

- 计算每一可能的组合，逐一比较准则函数。
- 适用于：d 或 $D-d$ 很小（组合数较少）的情况。

分支定界算法(Branch and bound, BAB)

- 从顶向下，有回溯(*a top-down approach, backtracking*)
- 应用条件：准则函数有单调性，即：

  对特征组 $\bar{X}_1 \supset \bar{X}_2 \supset \cdots \supset \bar{X}_i$ 有 $J(\bar{X}_1) \geq J(\bar{X}_2) \geq \cdots \geq J(\bar{X}_i)$

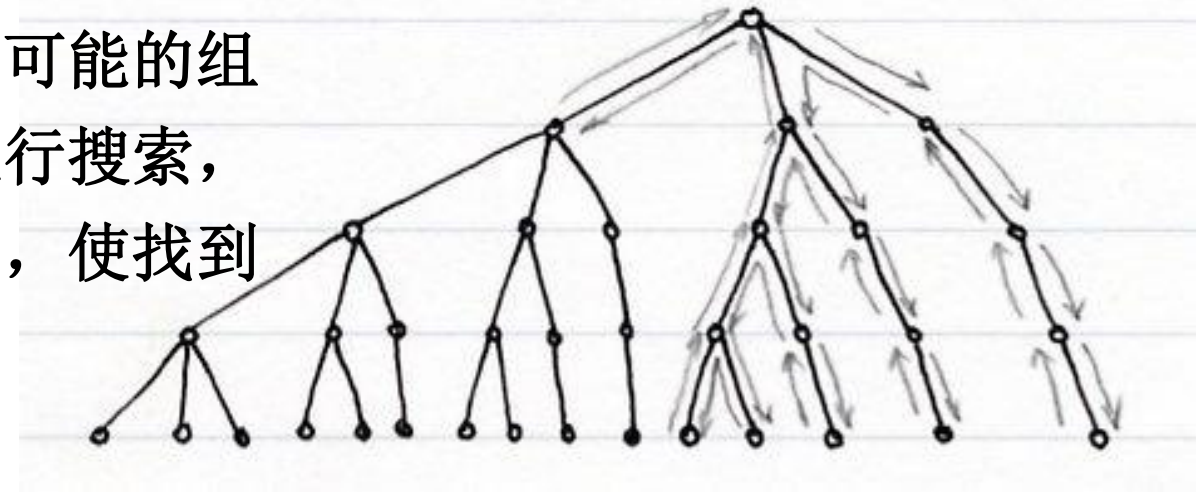
$$J(x_1, x_2, \cdots, x_d) \leq J(x_1, x_2, \cdots, x_d, x_{d+1})$$

J_1，J_2，J_3，J_4，J_5，J_C，J_B，J_D 都满足这一条件

5.2 特征选择的最优搜索方法(Optimal search method)

分支定界算法(BAB)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。



特点

① 最优搜索算法，所有可能的组合都被考虑到

② 前提：准则函数单调性（注：实际中可能不满足，因是估计值）

③ 节约计算与存储

④ $d \approx D/2$ 时最经济

5.2 特征选择的最优搜索方法(*Optimal search method*)

算法要点：

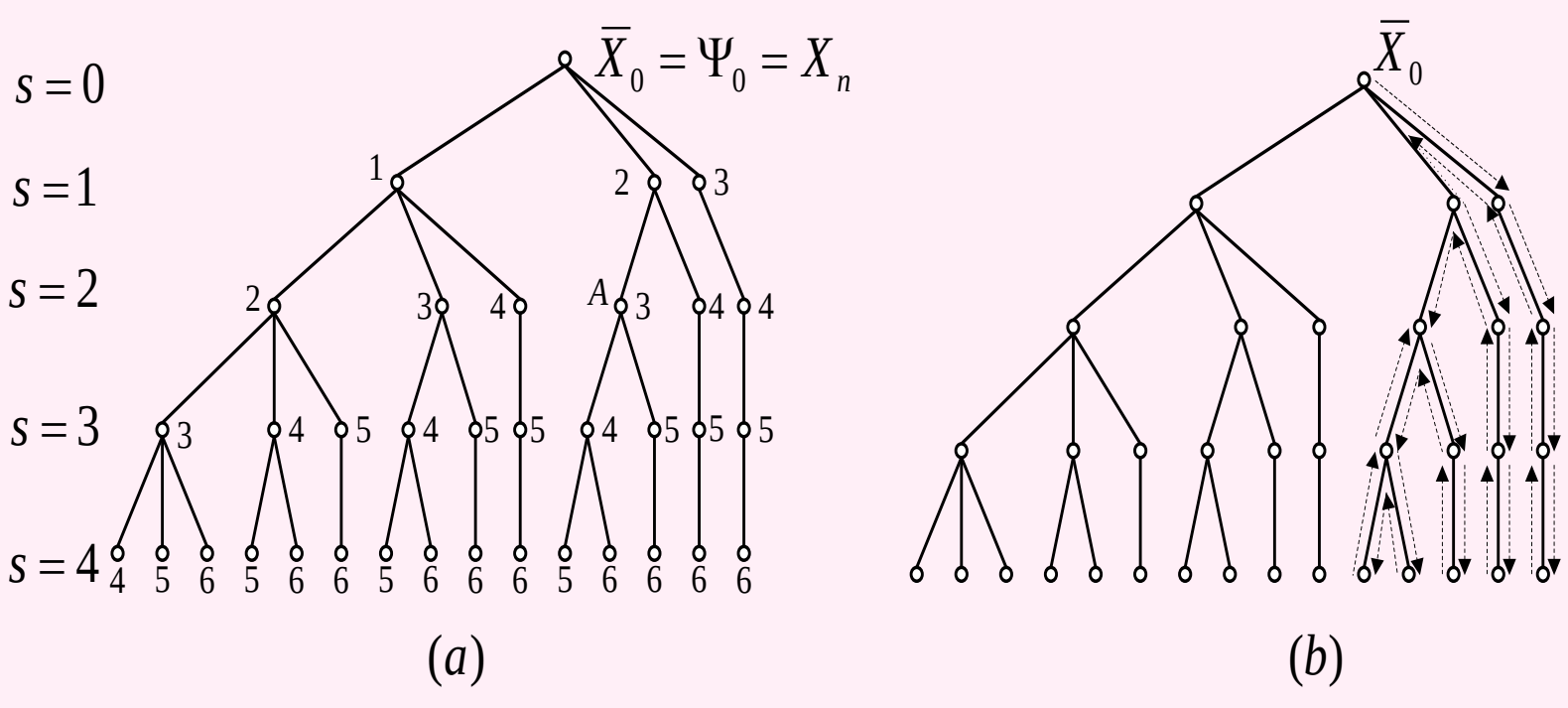
① 根结点为第0级，包含全体特征

② 每个结点上舍弃一个特征，各个叶结点代表选择的各种组合

③ 避免在整个树中出现相同组合的树枝和叶结点

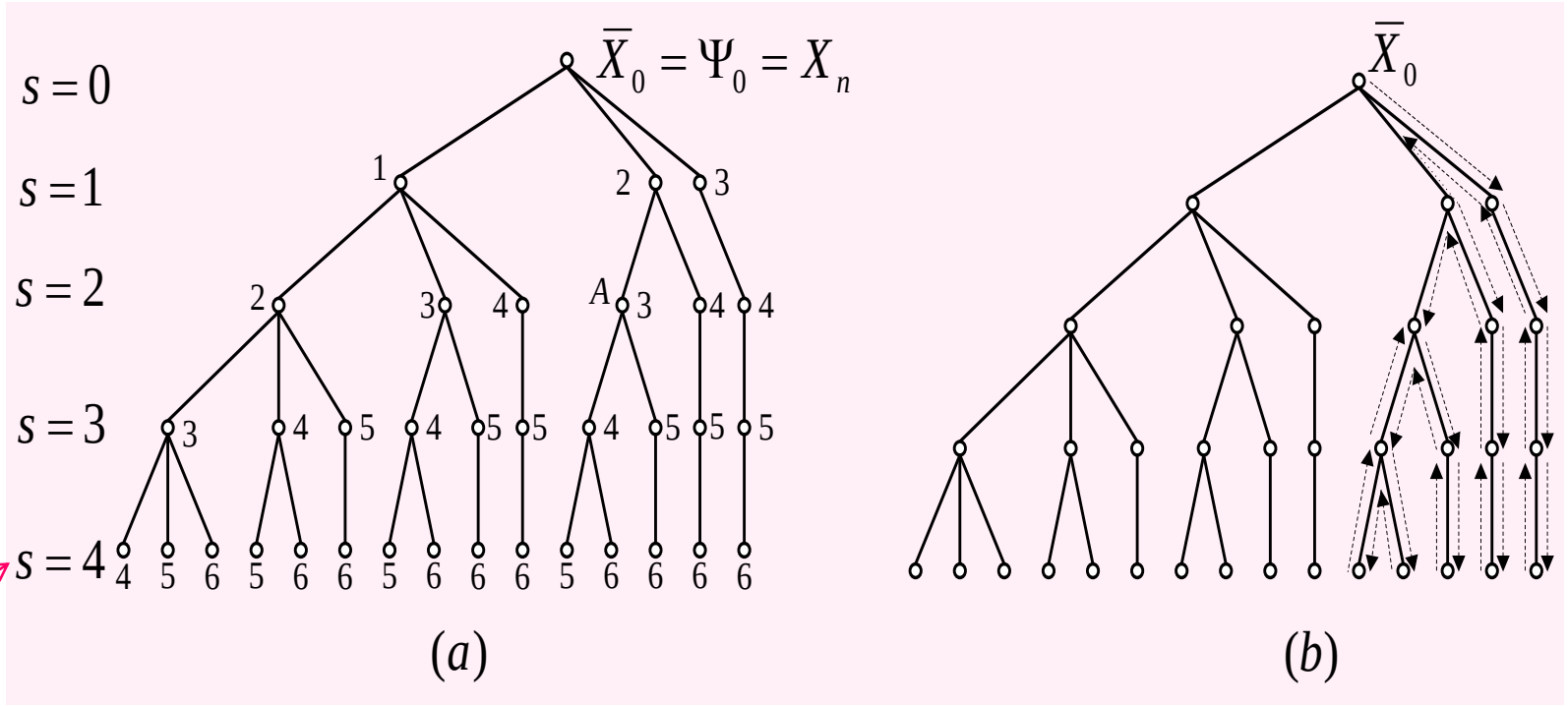
④ 记录当前搜索到的叶结点的最大准则函数值（界限B），初值置0

⑤ 从右侧开始搜索



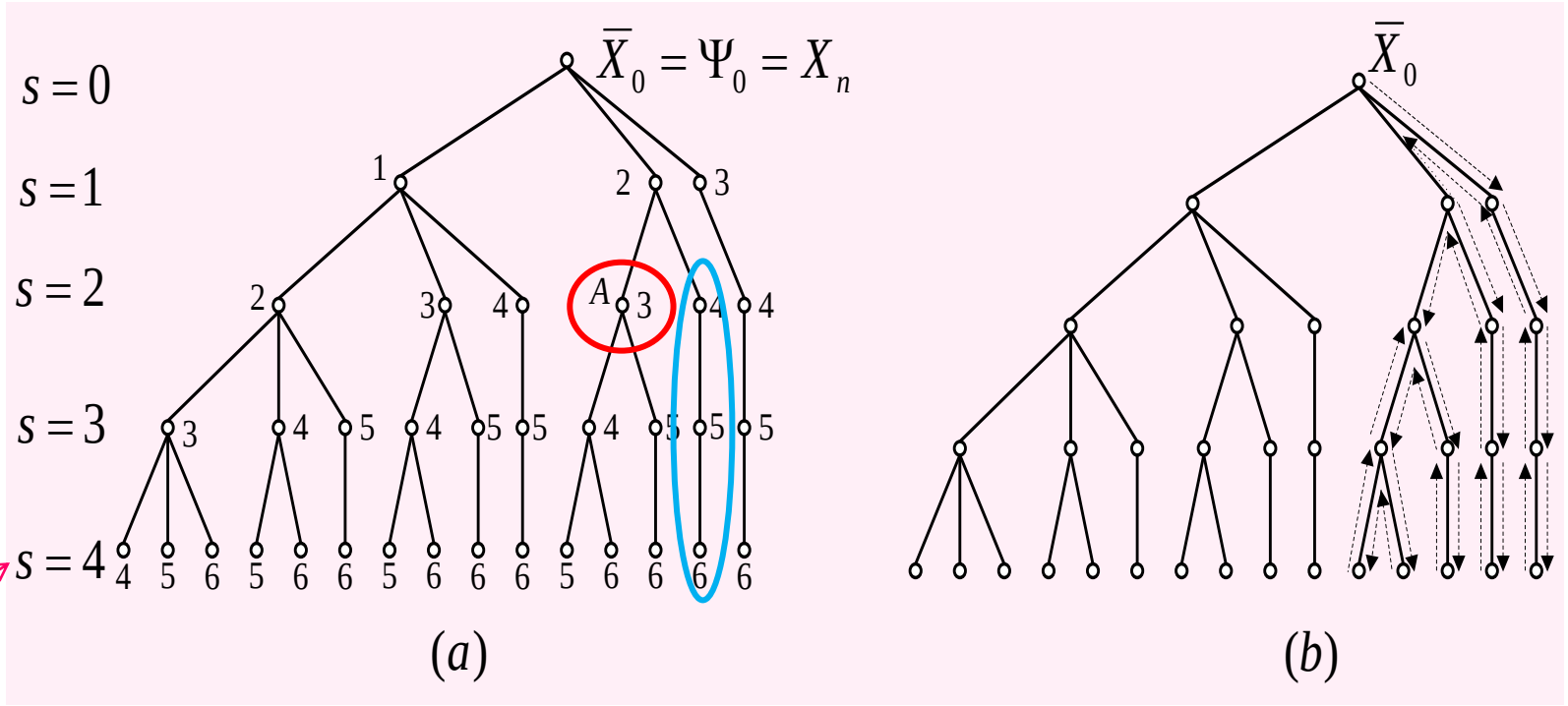
(a)搜索树(Search tree) (b)搜索回溯示意图(search with backtracking)

- 树的每个节点表示一种特征组合，树每一级的各节点表示从其父节点的特征组合中再去掉一个特征后的特征组合，其标号k表示去掉的特征是x_k。



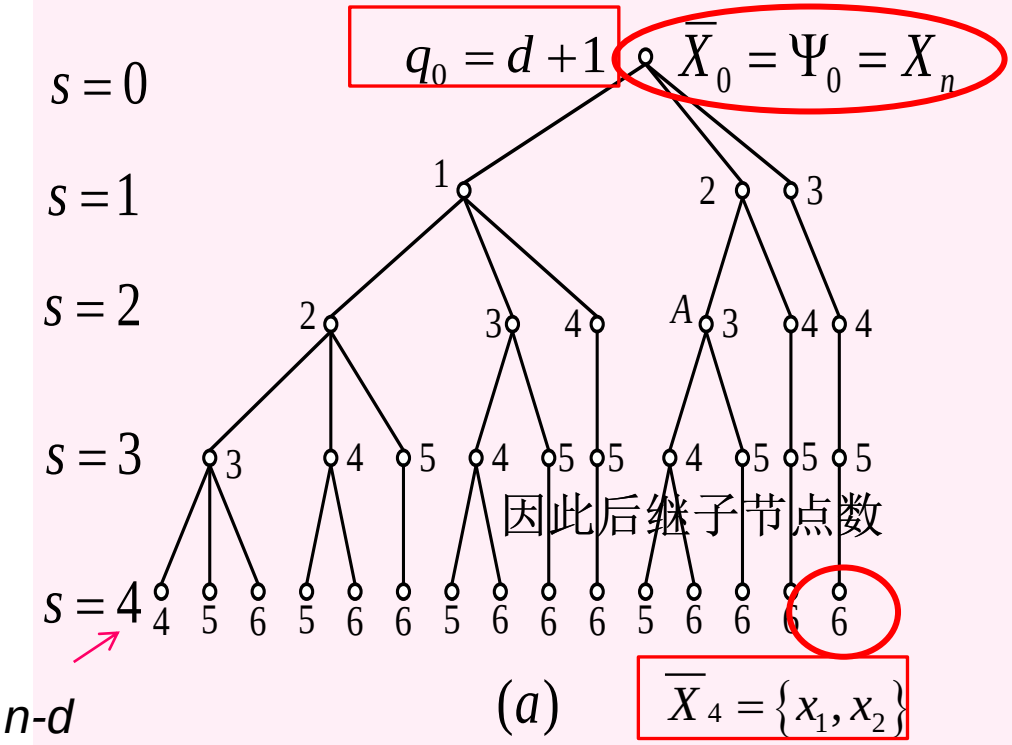
(a)搜索树(Search tree) (b)搜索回溯示意图(search with backtracking)

- 由于每一级只舍弃一个特征，因此整个搜索树除根节点的0级外，还需要*n-d* 级，即全树有*n-d* 级。在6个特征中选2个，故整个搜索树需要4级，第*n-d* 级是叶节点，共有叶节点 C_n^d 个。



(a)搜索树(Search tree) (b)搜索回溯示意图(search with backtracking)

- 对于任意某个子节点A而言,观察A及A右边的同父节点及其后的整个子树(要丢弃的特征标识),A节点要丢弃的特征不会出现在A点右边的兄弟子树上（要丢弃的特征标识）。



X_l 表示特征数目为 l 的特征集合。

$\overline{X}_s$ 表示舍弃 s 个特征后余下的特征集合。

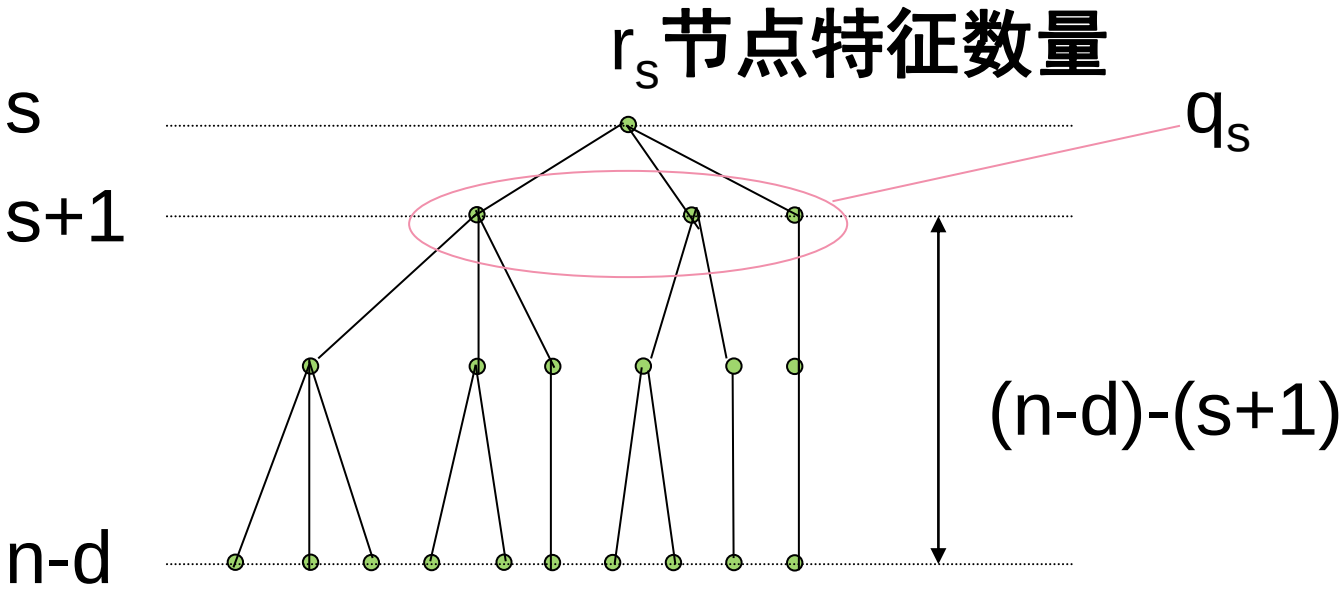
q_s 表示当前节点的子节点数。

Ψ_s 表示第 s 级当前节点上用来作为下一级可舍弃特征的特征集合。

r_s 表示集合 Ψ_s 中元素的数目。

- 由于从根节点要经历n-d级才能到达叶节点，s级某节点后继的每一个子节点分别舍弃ψ_s中互不相同的一个特征。
- 除了从树的纵的方向上一级丢弃一个特征，实际上从树的横的方向上，一个分支也轮换丢弃一个特征。

因此后继子节点数 $q_s = r_s - (n-d-s-1)$

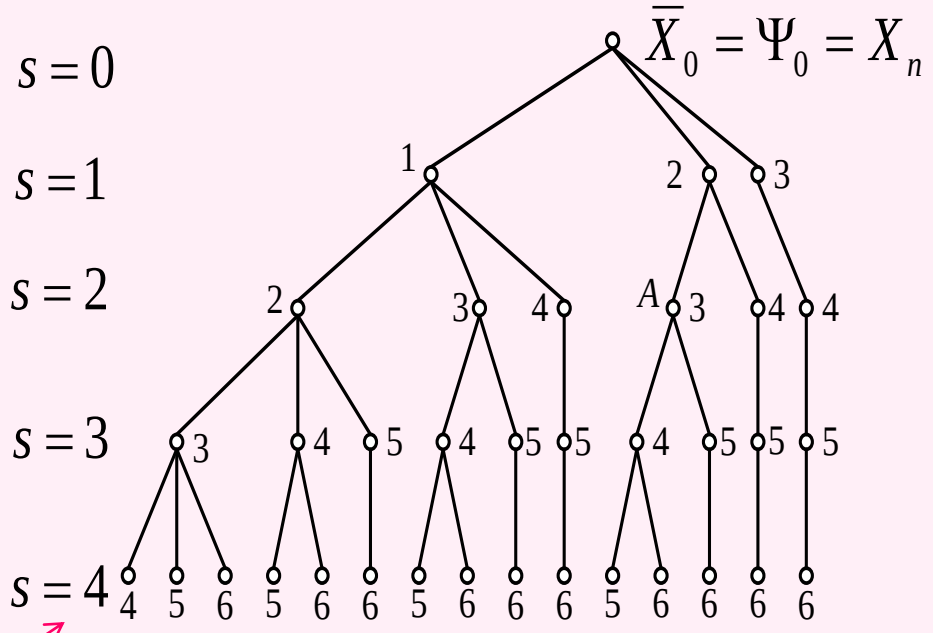


$$q_s = r_s - (n-d-s-1)$$

$$r_s = q_s + (n-d) - (s+1)$$

$$r_0 = n$$

$$q_0 = d+1$$



X_l 表示特征数目为 l 的特征集合。

$\bar{X}_s$ 表示舍弃 s 个特征后余下的特征集合。

q_s 表示当前节点的子节点数。

Ψ_s 表示第 s 级当前节点上用来作为下一级可舍弃特征的特征集合。

r_s 表示集合 Ψ_s 中元素的数目。

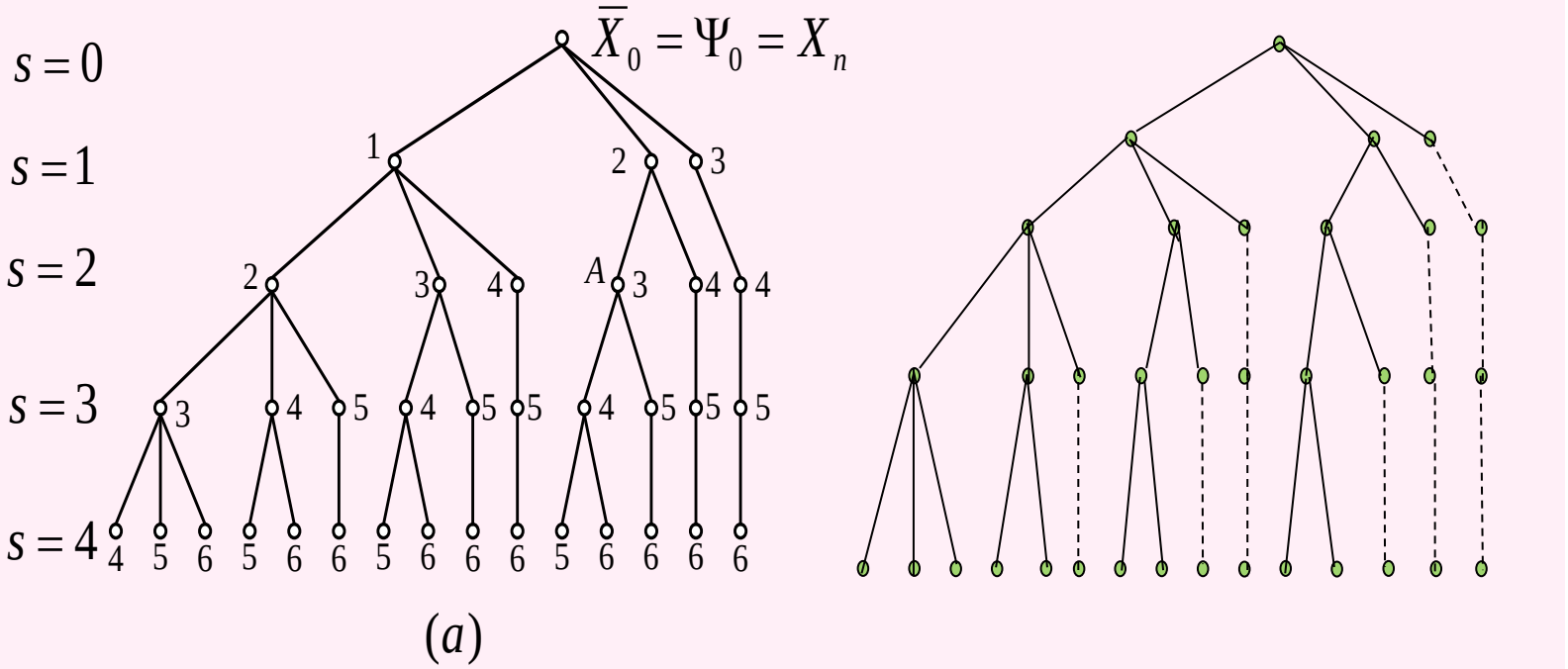
$$\Psi_0 = \{x_1, x_2, x_3, x_4, x_5, x_6\} \quad r_0 = 6$$

$$q_0 = 6 - (4 - 0 - 1) = 3$$

$$\Psi_1 = \{x_2, x_3, x_4, x_5, x_6\}$$

$$r_s = q_s + (n - d) - (s + 1)$$

$$q_s = r_s - (n - d - s - 1)$$



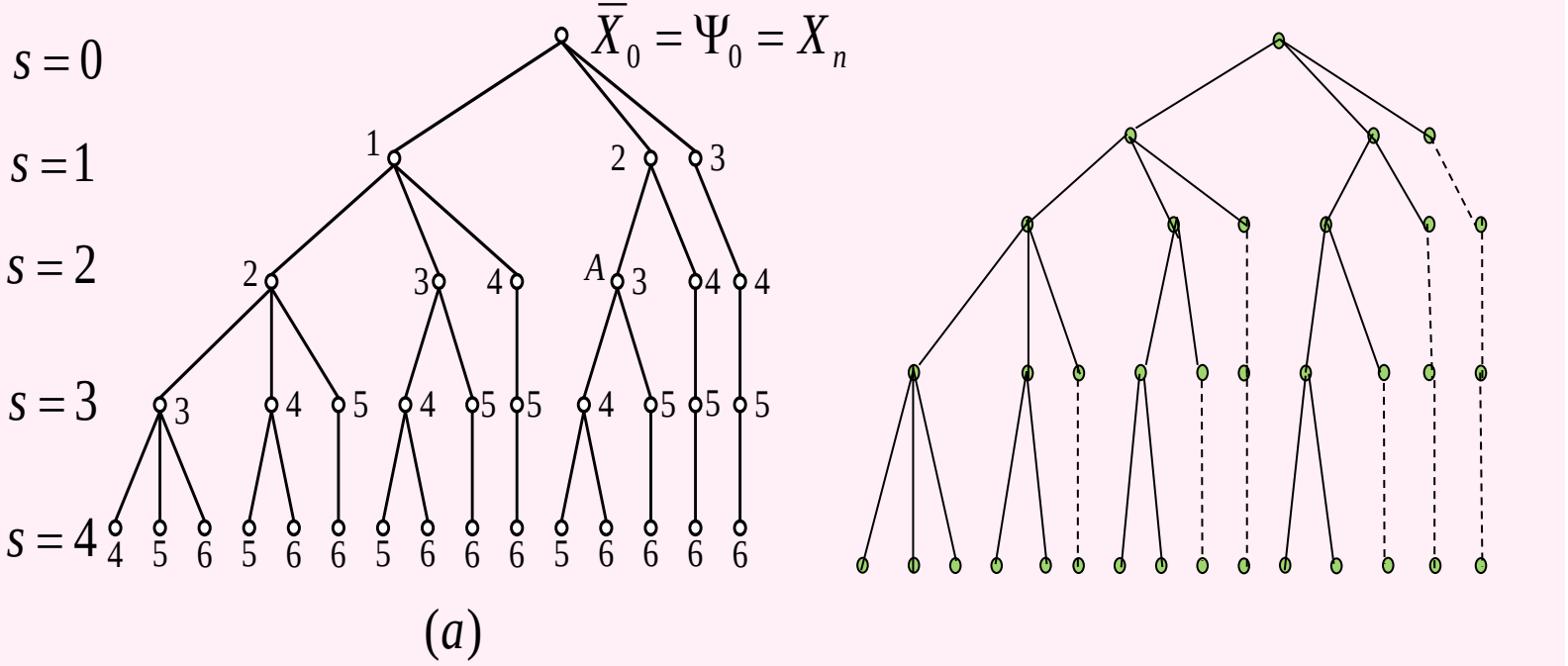
- 注意到每个节点都可以计算相应的J值。由于判据J值的单调性，使得：

$$J(x_1, x_2, \cdots, x_d) \le J(x_1, x_2, \cdots, x_d, x_{d+1})$$

上面的不等式表明，任何节点的J值均大于它所属的各子节点的J值。

5. 2 特征选择的最优搜索方法

BAB算法



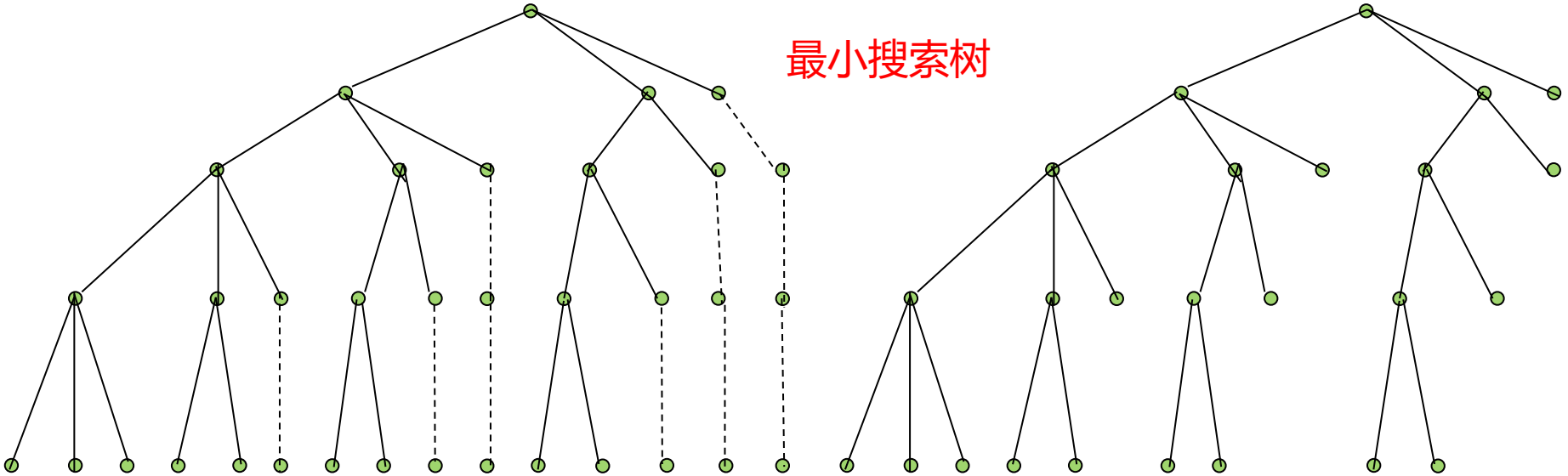
- 搜索过程是从上至下、从右至左进行

四个步骤：

1.向下搜索　　3.向上回溯

2.更新界值　　4.停止回溯再向下搜索

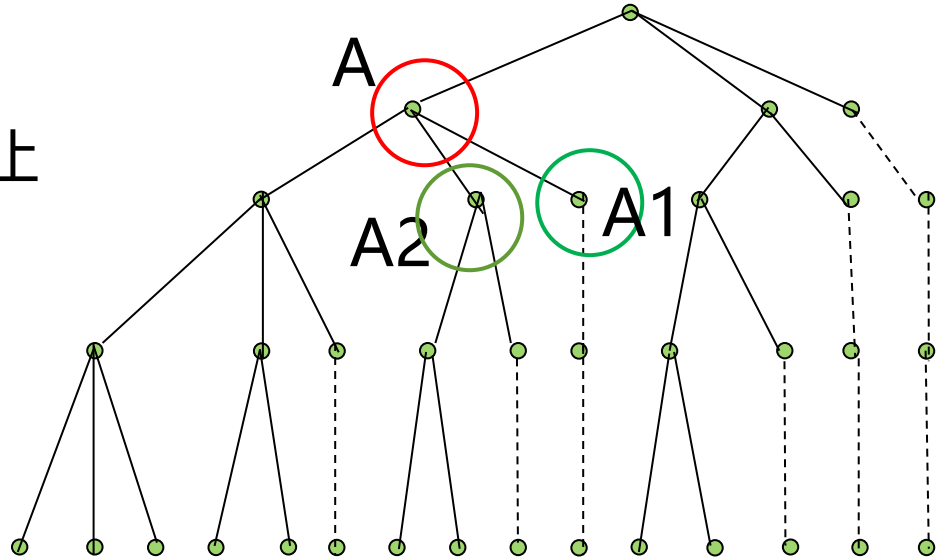
5.2 特征选择的最优搜索方法　　BAB算法



- 开始时置界值B=0，从树的根节点沿最右边的一支自上而下搜索。
- 对于一个节点，它的子树最右边的一支总是无分支的，即是1度节点或0节点（叶节点）。此时可直接到达叶节点，计算该叶节点的J值，并更新界值B。
- 图中的虚线可省略而得到最小搜索树。

- 向上回溯和停止回溯

  回溯到有分支的那个节点则停止回溯转入向下搜索。



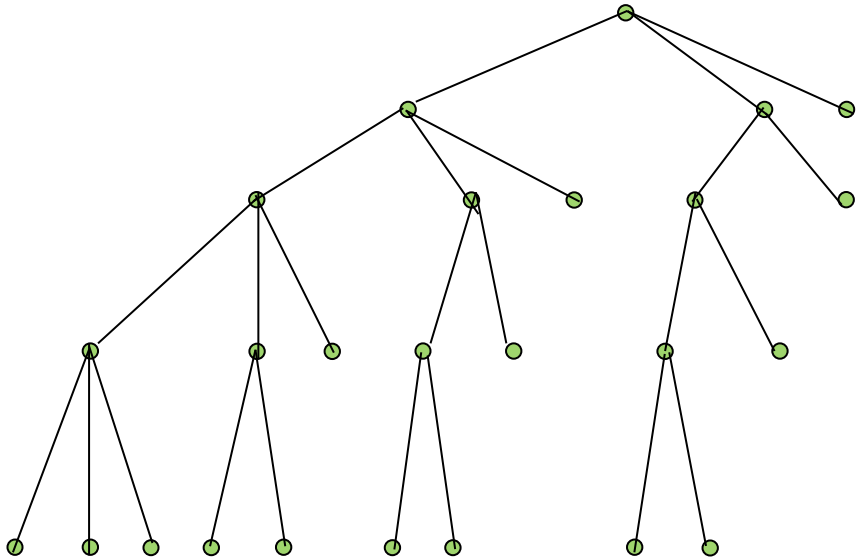
例如回溯到$q_{s-1}>1$ 的那个节点，则转入s深度的左边的最近的那个节点，使该节点成为当前节点，按前面的方法沿它最右边的子树继续搜索。

- 在搜索过程中先要判一下该节点的J值是否比B值大。若不大于B值，该节点以下的各子节点J值均不会比B大，故无需对该子树继续进行搜索。

5.2 特征选择的最优搜索方法

BAB算法

- 如果搜索到叶节点，且该叶节点代表的特征的可分性判据J值大于B，则更新界值，即B=J；否则不更新界值。



- 显然到达叶节点后，要向上回溯。重复上述过程，一直进行到J值不大于当前界值B为止。而对应的最大界值B的叶节点对应的d个特征组合就是所求的最优的选择。

5.2 特征选择的最优搜索方法

该算法的高效性能原因在于如下三个方面：

- 在构造搜索树时，同一父节点的各子节点为根的各子树右边的要比左边的少，即树的结构右边比左边简单；
- 在同一级中按最大的 J 值 ($J(X_l)$) 挑选舍弃的特征，树的节点的 J 值是左小右大，而搜索过程是从右至左进行的；
- 因 J 的单调性，树上某节点如 A 的可分性判据值 $J_A \leq B$，则 A 的子树上各节点的 J 值都不会大于 B，因此该子树各节点都可以不去搜索。

可知，有很多的特征组合不需计算仍能求得全局最优解。

5.2 特征选择的最优搜索方法

效果举例：比如在某组数据的实际中

D	d	穷举 (Exhaustive Search)	分支定界 (BAB)
12	4	$\binom{12}{4}$ = 495	42
24	12	$\binom{24}{12}$ = 2,704,156	13,369

5.3 非最优搜索方法(Heuristic启发式搜索)

非最优，但某些情况下最优，实现简单

（1）单独最优组合（Rank Search）

选前***d***个单独最佳的特征。

（2）**SFS法**（Sequential Forward Selection：顺序前进，前向贯序）

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

$$J(X_k + x_1) \geq J(X_k + x_2) \geq \cdots \geq J(X_k + x_{n-k})$$

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**SFS法**（GSFS, Generalized Sequential Forward Selection **)**

从底向上，每次增加 ***l*** 个特征。考虑了新增特征中的相关性

特点：计算量比**SFS**大，若 ***l*** **=*d***，（一步加满），则就是穷举法

（4）**SBS法**(Sequential Backward Selection 顺序后退，后向贯序)

从顶向下，每次减一个特征，与**SFS**相对，一旦失去，无法换回

$$J(\overline{X}_k - x_1) \geq J(\overline{X}_k - x_2) \geq \cdots \geq J(\overline{X}_k - x_{n-k})$$

15.3 非最优搜索方法(Heuristic启发式搜索)

（5）广义SBS法（GSBS）

从顶向下，每次减 r 个特征

（6）L-R法（增 l 减 r ，Plus-L Minus-R Selection ）

自底向上，每次增 l 个再减 r 个特征（$l > r$）

或向顶向下，每次减 r 个再增 l 个特征（$l < r$）

特点：带有局部回溯过程

（7）广义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。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

遗传算法是一种随机搜索算法（*Random Search*），通过模拟生物进化的现象（进化计算），把优化问题比作在无数可能的重组和突变组合中发现适应性最强的组合的问题。

- 用遗传算法进行特征选择（**D**个特征中选**d**个）
- 染色体（Chromosome）编码：*1为选择的特征，0为剔除的*

  二值字符串**m**，如 **(0 1 0 0 1 … 0 1 1 0 1 0 1)**，共 C_D^d种
- 适应度（fitness）函数：$f(m)$ 如某种类别可分性判据
- 选择概率模型：$p(f(m))$
- 基本操作：选择(Selection)，交叉（Crossover），变异(Mutation)

5.4 特征选择的遗传算法（*Genetic Algorithm*）

算法：

① 初始化，t=0，随机地产生一个包含L个染色体的种群 $M(0)$ ；

② 计算当前种群 $M(t)$ 中每一条染色体的适应度 $f(m)$ ；

③ 按照选择概率 $p(f(m))$ 对种群中的染色体进行选择，由选择出的染色体经过交叉、变异繁殖下一代染色体，组成下一代的种群 $M(t+1)$

④ 回到2，直到达到终止条件，输出适应度最大的染色体作为找到的最优解。终止条件通常是某条染色体的适应度达到设定的阈值。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

改进的遗传算法(*改进遗传算法和支持向量机的特征选择算法【杜卓明，冯静】*)

一个完整的遗传算法主要包括几个步骤：基因编码，种群初始化，选择操作，交叉操作，变异操作，结束条件判断等。

- **基因编码(coding)**

将选择的特征组合用一个{0，1}二进制串表示，0表示不选择对应的特征，1表示选择对应的特征。对惩罚参数C和核参数σ也采用二进制编码，根据范围和精度计算所需要的二进制串长度分别为l_c,l_σ。

- **种群初始化(population initialization)**

以a个特征中选取b个特征为例，确保在前a位二进制串中1出现的概率一定是b/a，两个参数部分的二进制码随机生成，染色体二进制长度为$l_a+l_c+l_\sigma$；然后以一定的种群规模进行种群初始化。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

- 选择操作(selection)

计算个体适应度(fitness)，即先对个体进行解码(decoding)，再用训练和测试样本计算SVM的正确分类率：

$$fitness = W_A \times SVM_{acuracy} + W_F \times \left(\sum_{i=1}^{l_a} C_i F_i \right)^{-1}$$

- W_A：SVM分类准确率权重，一般设置为75-100%
- $SVM_{accuracy}$：SVM分类准确率
- W_F：选择特征和惩罚参数乘积和逆的权重
- C_i：特征i的损失，如果没有关于损失的信息，可以设置为1
- F_i：1代表选择了特征i；0表示没有选择特征i。

然后采用轮盘赌选择法(Roulette Wheel Selection)，随机从种群中挑选一定的数目个体(individual)，再将适应度最好的个体作为父体，这个过程重复进行直到完成所有个体的选择。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

轮盘赌选择(Roulette Wheel Selection)又称比例选择算子，其基本思想是：各个个体被选中的概率与其适应度函数值大小成正比。设群体大小为N，个体x_i的适应度为$f(x_i)$，则个体x_i的选择概率为：

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

轮盘赌选择法可用如下过程模拟来实现：

(1)在［0，1］内产生一个均匀分布的随机数r。

(2)若$r \leqslant q_1$，则染色体x_1被选中。

(3)若$q_{k-1}<r \leqslant q_k$ （$2 \leqslant k \leqslant N$），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：

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

5.4 特征选择的遗传算法（*Genetic Algorithm*）

- 轮盘赌选择方法的实现步骤：
  （1）计算群体中所有个体的适应度值；
  （2）计算每个个体的选择概率；
  （3）计算积累概率；
  （4）采用模拟赌盘操作（即生成0到1之间的随机数与每个个体遗传到下一代群体的概率进行匹配）来确定各个个体是否遗传（复制，reproduction）到下一代群体中。

交叉操作(crossover)

由于交叉操作的随机性，会改变前a位二进制串中的1出现的概率，使其不等于b/a，这将导致不同个体特征矢量的维数不尽相同，所以进行以下操作。

首先将二进制编码分成两部分，前1_a位特征编码部分和后1_c+1_σ位参数编码部分。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

- 如下图所示

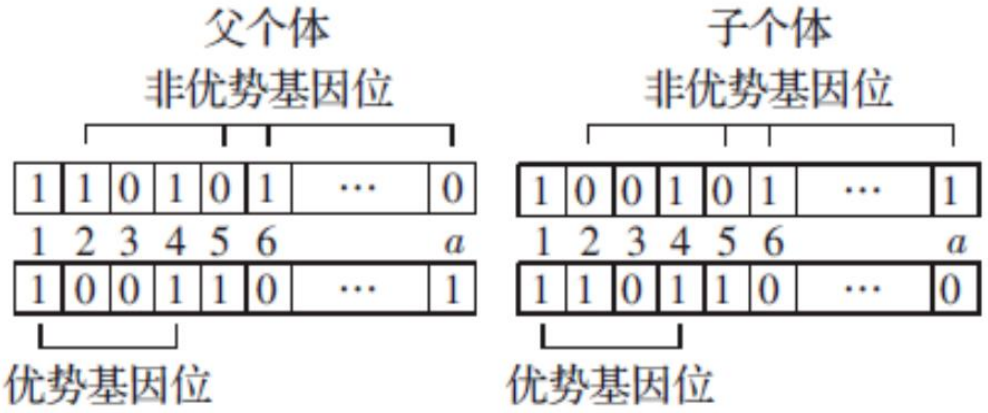


图 1　特征编码交叉操作

首先对比两个父体，找出两父体个体同为1的基因位，称之为“优势基因位”，例如第1，4位。然后找两父体其中一个为1的基因位，称之为“非优势基因位”，例如2，5，6，a。如果两父体中存在“优势基因位”，表明两父体对该基因位所对应的特征分量的选择意见趋于一致，该特征应在子代中予以保留。如果父代个体中存在“非优势基因位”，表明两个体在该特征上存在分歧。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

如果两父体个体存在e个“优势基因位”，则在子代中保留这些基因位，在“非优势基因位”中随机选择b-e个基因位，并令这些基因位为1，产生两个新个体。图1中两个子个体保留了第1、4位，子个体1在第2、5、6、a中随机选择了第6、a位并令其成为1，子个体2第2、5、6、a位中随机选择了第2、5位并令其为1。这样保证了子个体与父体选择的特征个数为b。

- 变异操作(mutation)

如果对特征编码进行翻转变异操作，那么将使二进制串中的为1的基因位发生变化，如果某一位由0变成1，则选择的特征数变为d+1，反之变为d-1。为解决这个问题可以使用下面的方法。

5.4 特征选择的遗传算法（*Genetic Algorithm*）

如图2，分别统计编码为1和0的基因位，分别在为1和0的基因位中随机选择一个二进制数，图2是第2和第5位相互交换，得到变异子个体。

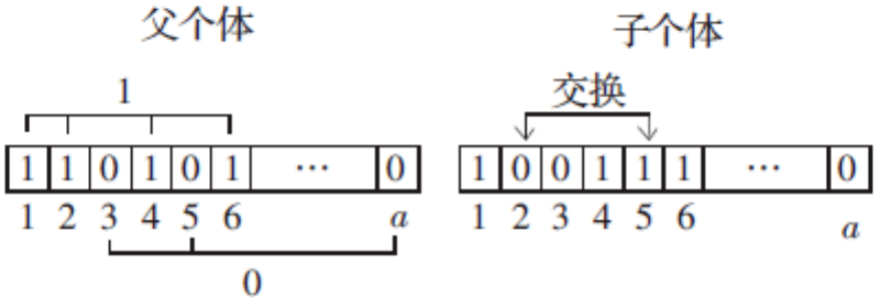


图2 特征编码变异操作

http://blog.csdn.net/li

❖ 结束条件（stopping criterion）

前面的选择，交叉，变异操作合起来称为遗传操作（genetic operators），当遗传操作到达设定的最大迭代次数时，算法结束。如果迭代遗传过程中，连续若干代最优个体不再变化，算法也可提前结束。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

Two-Step Procedure
(Filtering+Classification)

–Feature selection
(With some stand-alone criteria, like class-separation criteria or statistics criteria)

–Classification
(Using the selected features)

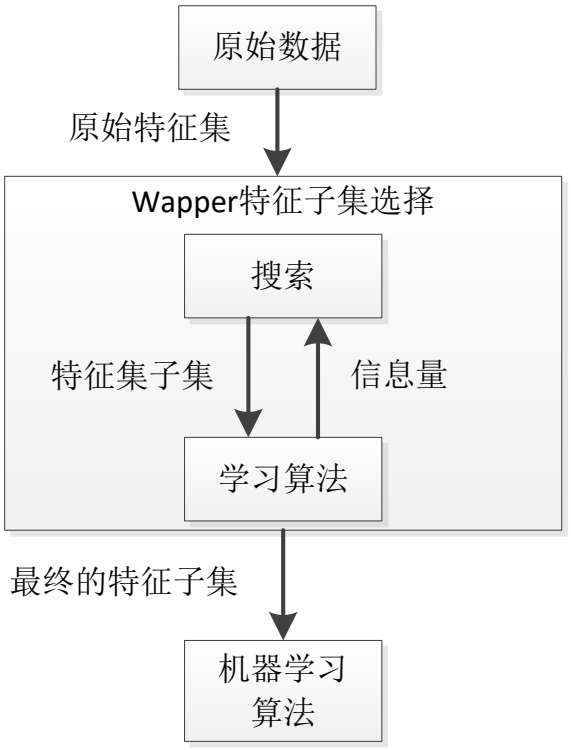
Recursive Procedure
(Wrapper approach)

–Classification (with all features)

–Feature selection (according to classification performance)

–Classification (using the selected features)

要求：分类器能处理特征维数很高的数据，样本有限时仍能得到较好的效果。



- 过滤式特征选择不考虑后续学习器。
- Wrapper方法直接把最终将要使用的模型的性能作为特征子集的评价标准， Wrapper方法的目的就是为给定的模型选择最有利于其性能的特征子集。
  从最终模型的性能来看， Wrapper方法比过滤式特征选择更好，但需要多次训练模型，因此计算开销较大。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

例如: **R-SVM(递归SVM)和 SVM-RFE（SVM递归特征剔除）**

1° 用当前所有候选特征训练线性支持向量机；

2° 评估当前所有特征在支持向量机中的相对贡献，按照相对贡献大小排序；

3° 根据事先确定的递归选择特征的数目选择出的排序在前面的特征，用这组特征构成新的候选特征，转1°，直到达到所规定的特征选择数目。

支持向量机的输出函数： $f(x) = w \cdot x + b$

评估特征在分类器中的贡献

线性核情况： **R-SVM:** $s_j = w_j\left(m_j^+ - m_j^-\right), \quad j = 1, \cdots, d$

SVM-RFE: $s_j^{RFE} = w_j^2$

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

Restrict training examples to good feature indices

$X = X_0(:, s)$

Train the classifier

1 $\alpha = SVM\text{-}train(X, y)$

Compute the weight vector of dimension length(**s**)

2 $w = \sum_k \alpha_k y_k x_k$ 计算权重向量

Compute the ranking criteria

3 $c_i = (w_i)^2, \quad \text{for all } i$ 计算排名标准

Find the feature with smallest ranking criterion

$f = argmin(c)$ 找到具有最小排序标准的特征

Update feature ranked list

$r = [s(f), r]$ 更新特征排名列表

Eliminate the feature with smallest ranking criterion

$s = s(1:f-1, f+1:length(s))$ 用最小排序准则消除特征

Output:

Feature ranked list **r**.

SVM-RFE的思想是根据SVM在训练时生成的权向量w来构造特征排序系数，每次迭代去掉一个排序系数最小的特征，最终得到所有特征属性的递减排序。

α代表分类器（实际上为svm正则项的拉格朗日算子向量）；2中K是训练样本数；Xk就代表了第k个样本的特征向量；W是输入空间特征权重向量；wi就是第i维特征的权重。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

回顾SVM

设训练集$\{(x_i, y_i)\}_{i=1}^N$，其中$x_i \in R^D, y_i \in \{+1, -1\}$，$x_i$为第$i$个样本，N为样本量，D为样本特征数。SVM寻找最优的分类超平面$\omega \cdot x + b = 0$。

SVM需要求解的优化问题为：

$$\begin{aligned} min \quad & \frac{1}{2}||\omega||^2 + C\Sigma_{i=1}^N \xi_i \\ s.t. \quad & y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, i = 1, 2, ..., N \\ & \xi_i \geq 0, i = 1, 2, ..., N \end{aligned}$$

而原始问题可以转化为对偶问题：

$$\begin{aligned} min \quad & \frac{1}{2}\Sigma_{i=1}^N \Sigma_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \Sigma_{i=1}^N \alpha_i \\ s.t. \quad & \Sigma_{i=1}^N y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, 2, ..., N \end{aligned}$$

其中，α_i为拉格朗日乘子。

最后ω的解为：$\omega = \Sigma_{i=1}^N \alpha_i y_i x_i$

两分类的SVM-RFE算法

SVM-RFE是一个基于SVM的最大间隔原理的序列后向选择算法。而特征k的排序准则得分定义为：

$$c_k = w_k^2$$

1)初始化原始特征集合S={1，2，…,D}，特征排序集R=[]

2)循环以下过程直至S=[]

获取带候选特征集合的训练样本；

用式$min \quad \frac{1}{2}\Sigma_{i=1}^N \Sigma_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \Sigma_{i=1}^N \alpha_i$训练SVM分类器，得到$\omega$；

用式$c_k = w_k^2, k = 1, 2, ..., |S|$计算排序准则得分；

找出排序得分最小的特征$p = arg \quad min_k c_k$；

更新特征集R=[p,R]；

在S中去除此特征：S=S/p。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

实例：

下面进行实验对比，数据集用公式产生：

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + e$$

x1到x5是由单变量分布生成的，e~N(0,1)。另外，原始的数据集中含有5个噪声变量 x6,…,x10与y无关，4个额外的变量x11,…x14，分别是x1,…,x4的关联变量，通过f(x)=x+N(0,0.01) 生成，这将产生大于0.999的关联系数。这样生成的数据能够体现出不同的特征排序方法应对关联特征时的表现。

在上述数据上运行所有的特征选择方法，并且将每种方法给出的得分进行归一化，让取值都落在0-1之间。对于RFE来说，由于它给出的是顺序而不是得分，我们将最好的5个的得分定为1，其他的特征的得分均匀的分布在0-1之间。

产生14个特征变量，一共750个样本

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

	X1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
Correlation	0.3	0.44	0	1	0.1	0	0.01	0.02	0.01	0	0.29	0.44	0	0.99
MIC	0.39	0.61	0.34	1	0.2	0	0.07	0.05	0.09	0.04	0.43	0.71	0.23	1
LinearRegression	1	0.56	0.5	0.57	0.27	0.02	0	0.03	0	0.01	0.6	0.14	0.48	0
Lasso	0.79	0.83	0	1	0.51	0	0	0	0	0	0	0	0	0.16
Ridge	0.77	0.75	0.05	1	0.88	0.05	0.01	0.09	0	0.01	0.59	0.68	0.02	0.95
RandomForestRegressor	0.37	0.54	0.1	0.64	0.25	0	0.01	0.01	0	0	0.61	0.53	0.09	1
Stability	0.65	0.67	0	1	0.63	0	0	0	0	0	0.32	0.44	0	0.51
RFE_lr	1	1	1	1	0.78	0.44	0	0.56	0.11	0.33	1	0.67	0.89	0.22
mean_score	0.66	0.68	0.25	0.9	0.45	0.06	0.01	0.1	0.03	0.05	0.48	0.45	0.21	0.6

Correlation：相关性；MIC最大信息系数算法；线性回归和正则化；lasso回归；Ridge岭回归；随机森林选择。稳定特征选择(Stability Selection): 用随机lasso算法的结果实现稳定特征选择；递归特征消除(Recursive Feature Elimination): 普通线性回归(lr)实现递归特征消除。最后，根据前面每个特征选择的方式的得到每个特征xi的平均得分。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

	X1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
Correlation	0.3	0.44	0	1	0.1	0	0.01	0.02	0.01	0	0.29	0.44	0	0.99
MIC	0.39	0.61	0.34	1	0.2	0	0.07	0.05	0.09	0.04	0.43	0.71	0.23	1
LinearRegression	1	0.56	0.5	0.57	0.27	0.02	0	0.03	0	0.01	0.6	0.14	0.48	0
Lasso	0.79	0.83	0	1	0.51	0	0	0	0	0	0	0	0	0.16
Ridge	0.77	0.75	0.05	1	0.88	0.05	0.01	0.09	0	0.01	0.59	0.68	0.02	0.95
RandomForestRegressor	0.37	0.54	0.1	0.64	0.25	0	0.01	0.01	0	0	0.61	0.53	0.09	1
Stability	0.65	0.67	0	1	0.63	0	0	0	0	0	0.32	0.44	0	0.51
RFE_lr	1	1	1	1	0.78	0.44	0	0.56	0.11	0.33	1	0.67	0.89	0.22
mean_score	0.66	0.68	0.25	0.9	0.45	0.06	0.01	0.1	0.03	0.05	0.48	0.45	0.21	0.6

MIC对特征一视同仁，它能够找出x3和响应变量y之间的非线性关系。

自学

LinearRegression的结果：噪声特征x6,…,x10得分接近0和y之间几乎没有关系。

Lasso：能够挑出一些优质特征，同时让其他特征的系数趋于0。可用于减少特征数，但是对于数据理解来说不是很好用。例如在结果表中，x11,x12,x13的得分都是0，好像他们跟输出变量之间没有很强的联系，但实际上不是这样的。

5.5 以分类性能为准则的特征选择方法（Wrapper方法）

	X1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
Correlation	0.3	0.44	0	1	0.1	0	0.01	0.02	0.01	0	0.29	0.44	0	0.99
MIC	0.39	0.61	0.34	1	0.2	0	0.07	0.05	0.09	0.04	0.43	0.71	0.23	1
LinearRegression	1	0.56	0.5	0.57	0.27	0.02	0	0.03	0	0.01	0.6	0.14	0.48	0
Lasso	0.79	0.83	0	1	0.51	0	0	0	0	0	0	0	0	0.16
Ridge	0.77	0.75	0.05	1	0.88	0.05	0.01	0.09	0	0.01	0.59	0.68	0.02	0.95
RandomForestRegressor	0.37	0.54	0.1	0.64	0.25	0	0.01	0.01	0	0	0.61	0.53	0.09	1
Stability	0.65	0.67	0	1	0.63	0	0	0	0	0	0.32	0.44	0	0.51
RFE_lr	1	1	1	1	0.78	0.44	0	0.56	0.11	0.33	1	0.67	0.89	0.22
mean_score	0.66	0.68	0.25	0.9	0.45	0.06	0.01	0.1	0.03	0.05	0.48	0.45	0.21	0.6

自学

Ridge：将回归系数均匀的分摊到各个关联变量上，x11,…x14和x1,…,x4的得分非常接近。
随机森林：基于不纯度的排序结果非常鲜明，在得分最高的几个特征之后的特征，得分急剧下降。
稳定性选择：既能够有助于理解数据又能够挑出优质特征，在结果表中就能很好的看出。像Lasso一样，它能找到那些性能比较好的特征x1,x2,x4,x5，同时，与这些特征关联度很强的变量也得到了较高的得分。