§5.3 非最优搜索方法(Heuristic启发式搜索)

（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

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

（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 遗传算法

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。

改进遗传算法

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$\mathbf{l_c}$,$\mathbf{l_\sigma}$。

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$\mathbf{l_a}$+$\mathbf{l_c}$+$\mathbf{l_\sigma}$；然后以一定的种群规模进行种群初始化。

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：

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

- 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。

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≤q_1，则染色体x_1被选中。

(3)若q_{k-1}<r≤q_k （2≤k≤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：

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

❖ 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。

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。

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法(Wrapper)

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

Restrict training examples to good feature indices

$X=X_0(:,\mathbf{s})$

Train the classifier

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

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

2 $\mathbf{w}=\sum_k\alpha_k y_k\mathbf{x}_k$ 计算权重向量

Compute the ranking criteria

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

Find the feature with smallest ranking criterion

$f=argmin(\mathbf{c})$ 找到具有最小排序标准的特征

Update feature ranked list

$\mathbf{r}=[\mathbf{s}(f),\mathbf{r}]$ 更新特征排名列表

Eliminate the feature with smallest ranking criterion

$\mathbf{s}=\mathbf{s}(1:f-1,f+1:\text{length}(\mathbf{s}))$ 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。