# 特征选择

特征选择的一些算法介绍。

## 定义

1）理想化的定义：寻找一个对目标概念必要和充分的元素个数最小的特征子集。

2)经典的定义：从原始的N个特征中选出一个由M个特征组成的特征子集，其中M事先给定，M<=N，使得在原集合所有元素个数为M的子集中，该子集对于某种评价标准最优。

3)提高预测精度角度的定义：选择特征子集来增加分类精度，或者在不降低分类器精度的条件下降低特征集的维数

特征选择是机器学习领域的一个难题，本质上是一个组合优化问题，求解组合优化问题最直接的方法就是搜索，理论上可以通过穷举法来搜索所有可能的特征组合，但是n个特征的搜索空间为2n，因此穷举法难以实际应用。其他的搜索方法有启发式的搜索和随机搜索，这些搜索策略可以在运算效率和特征子集质量之间寻找到一个较好的平衡点。



特征选择的基本步骤

## 特征选择算法分类

按照搜索策略划分特征选择算法

1)全局最优策略

全局最优策略有穷举法和分支定界法。分支定界法要求指定所选子集中的特征个数，并要求评价函数满足单调性，的搜索空间是一棵树。

2)启发式搜索策略

* 单独最优策略：分别对每个特征单独进行评价，根据各特征的评价值对特征排序，删掉排在后面的特征。优点在于计算复杂度低，运行速度快。
* 序列前向选择方法(SFS)：即顺序前进法，是一种最简单的自下而上的搜索方法。先把所需要的特征集合初始化为一个空集，每次从入选的特征选择一个使得评价函数值最大的特征，加入已选特征集，直到特征数满足要求为止。SFS算法的运算量相对较小，其局限性在于特征之间的相关性没有得到充分的考虑，并且特征一旦加入，就无法剔除。（贪心算法）
* 广义序列前向选择方法(GSFS)：该方法是SFS算法的推广，每次从未入选的特征中选择l个特征，加入到己选特征集中，使得这l个特征加入后评价函数的值达到最大。由于每次添加需要计算多种特征组合的情况，计算量比SFS大。
* 序列后向选择方法(SBS)：即顺序后退法。先把所需要的特征集合初始化为一个包含所有特征的全集，然后在算法的每次迭代过程中删除一个特征，使得剩余的特征集合的评价函数值达到最优，直到剩余特征个数符合要求。SBS方法对特征集合的评价是在高维空间进行，因此计算量较大。SBS的优势在于考虑了特征之间的统计相关特性，实际表现通常要优于SFS。
* 广义序列后向选择方法(GSBS)：该方法是SBS算法的推广算法，根据预先定义好的评价函数，在算法的每次迭代过程中，一次性删除r个对评价函数值贡献较小的特征。
* 增l减r法：将SFS和SBS相结合，先使用SFS算法逐个选入l个最佳特征，然后使用SBS算法剔除r个最差特征，直到特征个数达到要求为止。（l>r）
* 浮动搜索方法：前面各种算法的l、r都是固定的，浮动搜索方法采用变化的步长，即在特征选择的过程中，添加的特征个数l和删除的特征个数r是变化的。

3）随机搜索策略

随机搜索策略的典型方法包括：遗传算法和模拟退火算法等。

* 遗传算法：借鉴生物的自然选择和遗传进化机制而开发出的一种全局优化自适应概率搜索算法，使用群体搜索技术，通过对当前群体进行选择、交叉、变异等一系列遗传操作，从而产生新一代的群体，并逐步使群体进化到包含或接近最优解的状态。

根据Jain等人的实验，浮动搜索算法和遗传算法是各种搜索算法中性能较好的算法。

## 遗传算法

基本过程：

1）**编码**：采用合适的编码方式进行编码，每个编码代表解空间的一个解。（将个体表示为符号串）

2）**初始群体**的产生：形成初始的群体数据。

3）**适应度计算**：以个体适应度的大小来评定各个个体的优劣程度，从而决定其遗传机会的大小。（适应度函数需要是非负的。）

4）**选择**：把当前群体中适应度较高的个体按某种规则或模型遗传到下一代群体中。

1. 计算群体中所有个体的适应度总和F

2. 计算每个个体的适应度f，则f/F可以作为遗传到下一代的概率。

3. 根据生成的概率分布生成随机数，决定每个个体被选中的次数。

5）**交叉运算**：交叉运算是遗传算法中产生新个体的主要操作过程，它以某一概率相互交换某 两个个体之间的部分染色体。以单点交叉的方法为例，具体操作过程是：  
       • 先对群体进行随机配对；  
       • 其次随机设置交叉点位置；  
       • 最后再相互交换配对染色体之间的部分基因。

6）**变异运算**：变异运算是对个体的某一个或某一些基因座上的基因值按某一较小的概率进

行改变，它也是产生新个体的一种操作方法。

## 模拟退火算法**(SA, Simulated Annealing)**

贪心算法的问题是每次选择的是当前的最优解，因此很可能最后得到局部最优解（鼠目寸光）。模拟退火算法**以一定的概率来接受一个比当前解要差的解**，因此有可能会跳出这个局部的最优解，达到全局的最优解。算法描述：

1）若J( Y(i+1) )>= J( Y(i) ) (即移动后得到更优解)，则总是接受该移动；

2）若J( Y(i+1) )< J( Y(i) ) (即移动后的解比当前解要差)，则以一定的概率接受移动，而且这个概率随着时间推移逐渐降低（逐渐降低才能趋向稳定）

　　这里的“一定的概率”的计算参考了金属冶炼的退火过程，这也是模拟退火算法名称的由来。模拟退火算法是一种随机算法，并不一定能找到全局的最优解，可以比较快的找到问题的近似最优解。 如果参数设置得当，模拟退火算法搜索效率比穷举法要高。

ref：

《机器学习中的特征选择算法研究》（姜百宁， 海洋大学硕士论文）

（遗传算法） <http://blog.csdn.net/b2b160/article/details/4680853/>

（退火算法） <http://www.cnblogs.com/heaad/archive/2010/12/20/1911614.html>