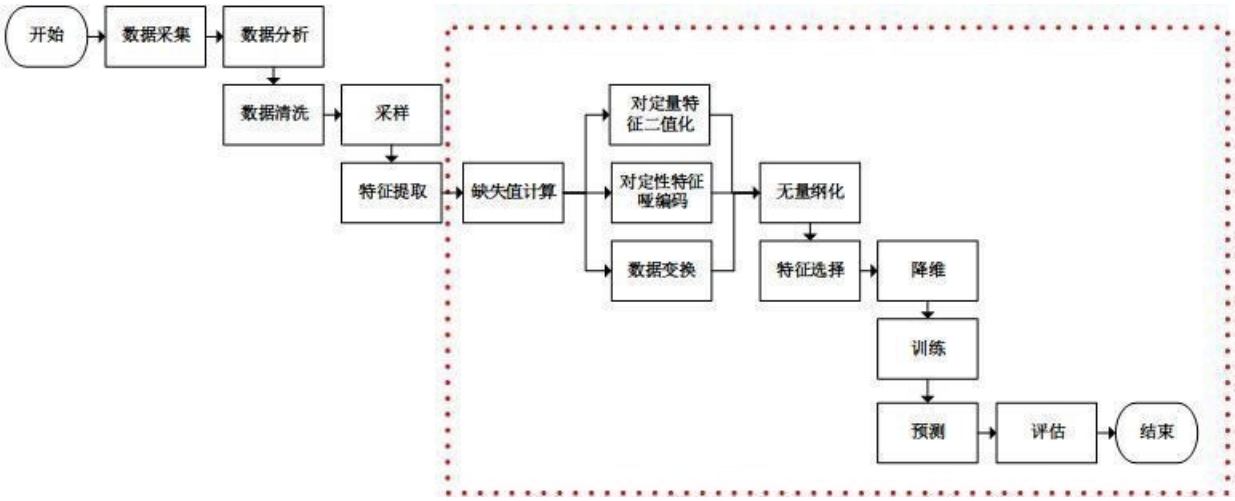
特征选择方法

1. General



上图为数据挖掘或机器学习基本的场景描述，当特征提取后进行预处理后，此时需要选择对训练模型有意义的特征，这个过程我们称为特征选择。

特征选择：也称特征子集选择（Feature Subset Selection , FSS），或属性选择（Attribute Selection）；是指从特征集选取一个特征子集，使构造出来的模型更好。

> 特征选择技术的常常用于许多特征但样本（即数据点）相对较少的领域。特征选择应用的典型用例包括：解析书面文本和微阵列数据，这些场景下特征成千上万，但样本只有几十到几百个

1.1 Why

在机器学习实际运用中，特征数量往往非常多，其中可能存在许多不相关的特征，特征之间也可能存在相互依赖，容易导致的后果：

- 训练时间变长
- 维度灾难，模型复杂
- 过拟合(决策树)

特征选择能剔除irrelevant或redundant的特征，从而达到减少特征个数、提升模型准确度、减少训练时间的目的

2. Procedure

通常来说，会从两个方面来考虑选择特征：

- 特征是否发散：如果样本在某特征上基本无差异，则该特征则对样本的区分无意义
- 特征与目标的相关性：与目标相关性高的特征，应该作为最优特征

特征选择一般过程如下：

1. **产生过程(Generation Procedure)**：从特征全集中选取中特征子集
2. **损失函数评价(Evaluation Function)**：用损失函数对该特征进行评价
3. **停止准则比较(Stopping Criterion)**：若评价结果比停止准则(一般为阈值)差，则重复1,2步骤，否则走4
4. **验证过程(Validation Procedure)**：对特征子集验证其有效性

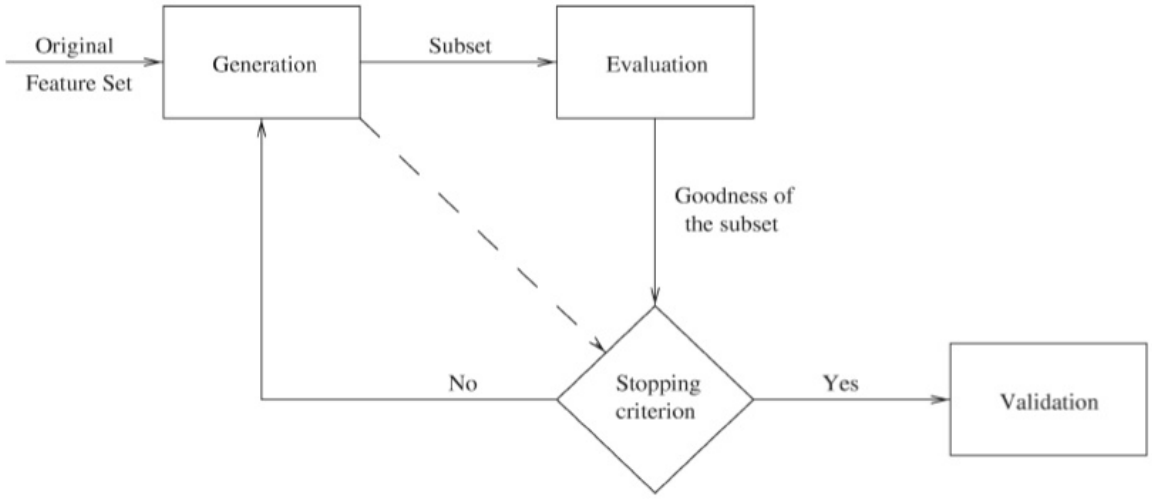


Fig. 1. Feature selection process with validation.

2.1 Generation Procedure

产生过程是搜索特征子空间的过程。如果特征全集包含N个特征，则要生成的候选特征子集的总数是2^N，即使是中等规模的N，这也是个巨大的数字。解决这一问题的算法有完全搜索(Complete)，启发式搜索(Heuristic)，随机搜索(Random)

2.1.1 Complete

完全搜索分为穷举搜索(Exhaustive)和非穷举搜索(Non-Exhaustive)两类

- 广度优先搜索(Breadth First Search)

  广度优先遍历特征子空间；枚举了所有特征组合，时间复杂度为$O(2^n)$，实用性不高

- 分支限界搜索(Branch and Bound)

  穷举搜索的基础上加上了分支限界；例如：剪掉某些不可能搜索出比当前最优解更优的分支。

- 定向搜索(Beam Search)

  首先选择N个得分最高的特征作为特征子集，将其加入一个限制最大长度的优先队列，每次从队列中取得得分最高的子集，然后穷举向该子集加入1个特征后产生的所有特征集，将这些特征集加入队列

- 最优优先搜索(Best First Search)

  在定向搜索的基础上不限制优先队列的长度

2.1.2 Heuristic

- 序列前向选择(SFS , Sequential Forward Selection)

  从空集开始，每次选择一个特征加入子集X，使得特征函数$F(X)$最优；

缺点：无法剔除特征

- 序列后向选择(SBS , Sequential Backward Selection)

  从特征全集开始，每次剔除一个特征，得到子集X，使得特征函数$F(X)$最优；

  缺点：无法加入剔除的特征

- 双向搜索(BDS , Bidirectional Selection)

  使用SFS和SBS同时构建子集X，直到两者的X相同时，停止搜索

- 增L去R选择算法(LRS , Plus-L Minus-R Selection)

  从空集开始，每次加入L个，减去R个，选最优（L>R)或者从全集开始，每次减去R个，增加L个，选最优(L<R)。

- 序列浮动选择(Sequential Floating Selection)

  在LRS的基础上取浮动的L和R参数

- 决策树(DTM , Decision Tree Method)

  决策树的剪枝过程。评价函数为信息增益

2.1.3 Random

- 随机产生序列选择算法(RGSS , Random Generation plus Sequential Selection)

  随机产生一个特征子集，然后使用SFS或SBS求解，最终求得各个子集的最优解

- 模拟退火算法(SA , Simulated Annealing)

  选定一个点，求得改点的解，并迭代之后点的解进行比较求得最优解；但在一定概率范围内可认为比最优解大的解继续向后迭代，最终取得近似全局最优解

- 遗传算法(GA , Genetic Algorithms)

  随机产生一组特征子集，并用评价函数进行评分，通过交叉、突变形式繁殖出下一代特征子集，经过N此淘汰后可得到评价函数最高的特征子集

随机算法缺点：依赖随机因素，有实验结果难以重现

2.2 Evaluation Function

评价函数的作用是评价产生过程所提供的特征子集的好坏。

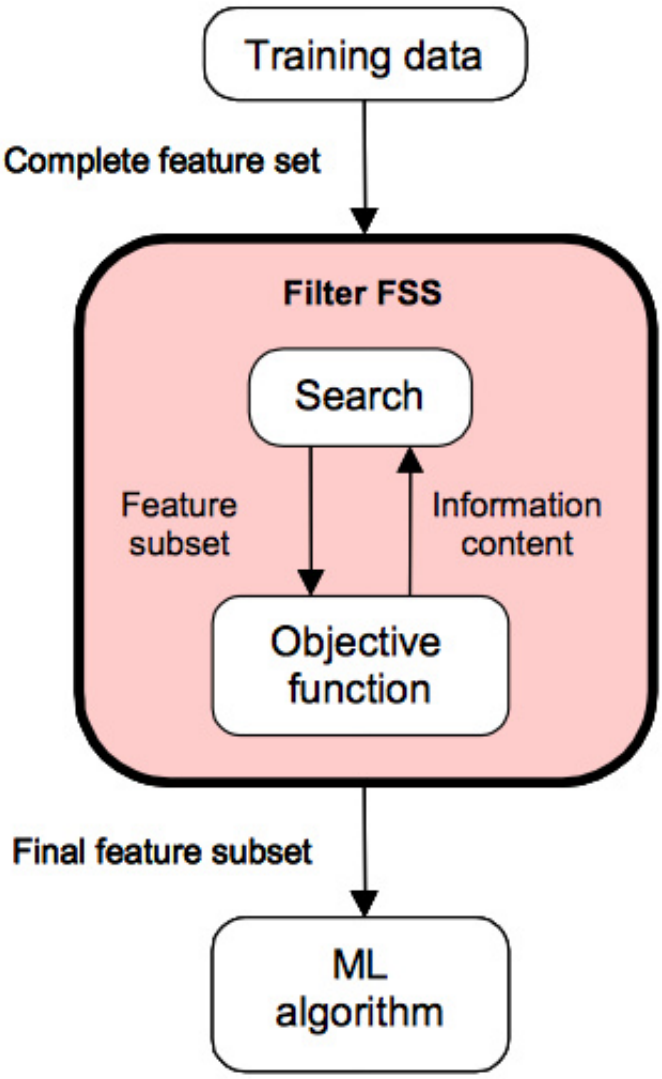
主要有三大类：

- Filter：过滤法
- Wrapper：包装法
- Embedded：集成法

2.2.1 Filter

过滤法通过分析特征子集内部的特点来衡量好坏。一般用作预处理

按照发散性或相关性对各个特征进行评分，设定阈值或者待选择特征的个数，选择特征



常用方法：

- 方差阈：计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征。默认情况下去除方差等于0的特征
- 相关系数法：计算各个特征对目标值的相关系数以及相关系数的P值；好的特征子集包含的特征应该与分类的相关度较高，而特征之间相关度较低
- 卡方检验：`只能用于二分类`，检验定性自变量对定性因变量的相关性
- 互信息法：也是检验定性自变量对定性因变量的相关性
- 一致性：若样本1与样本2属于不同的分类，但在特征A、B上的取值完全一样，那么特征子集{A,B}不应该选作最终的特征集
- 信息增益：ID3算法中的信息增益比较，取信息增益较高的特征子集

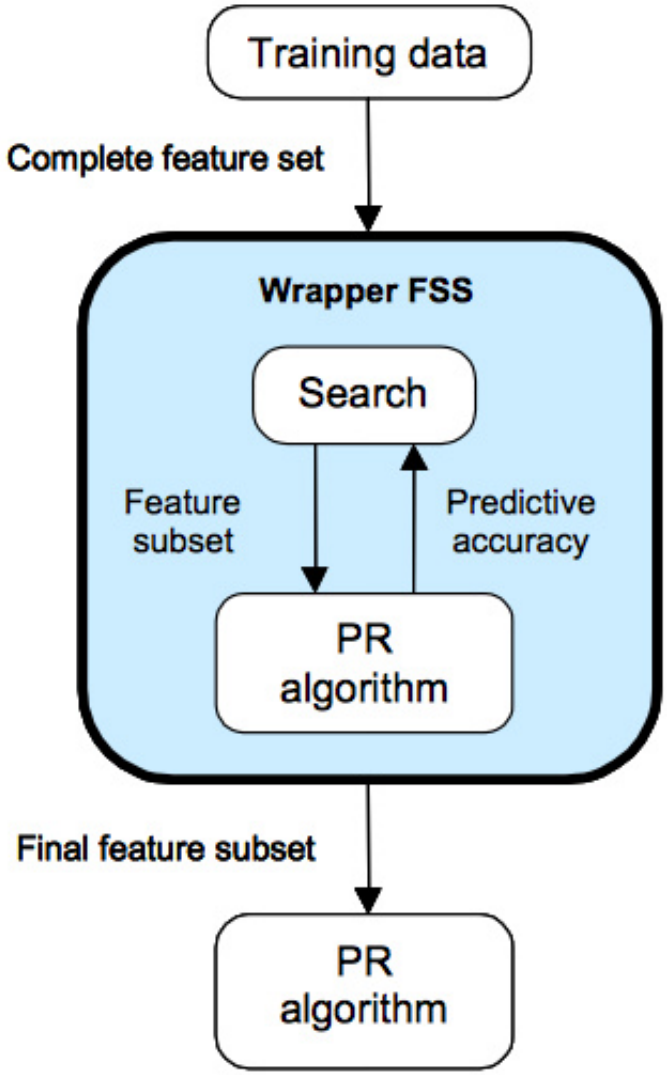
优缺点：

1. 优点：
    - 执行高效：通过涉及数据集上的非迭代运算
    - 通用性：评估数据的固有属性，而不是评估特定分类器的相互作用
2. 缺点：
    - 大型子集的倾向性：由于目标函数通常是单调的，所以其更倾向与选择全特征集作为最优

解，故不得不使得用户设定阈值中断

2.2.2 Wrapper

Objective Function是一个模式分类器，它通过统计重采样或交叉验证，预测准确度来评估特征子集



常用方法：

- 递归特征消除法：使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练

优缺点：

1. 优点：
    - 准确性：比Filter获得更好的识别率
    - 泛化能力：避免过拟合的机制
2. 缺点
    - 执行速度慢：必须为每个特征子集训练分类器
    - 缺乏通用性：与评价函数所使用的分类器的偏向有关

2.2.3 Embedded

集成法，先使用某些机器学习的算法或模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣

常用方法：

- 基于惩罚项的特征选择法：筛选特征、降维
- 基于树模型的特征选择法：GBDT

FYI

- 机器学习之特征工程-特征选择
- Ricardo Gutierrez-Osuna, Introduction to Pattern Analysis (LECTURE 11: Sequential Feature Selection)
- Feature Selection for Classification
- [特征选择常用算法综述](http://www.cnblogs.com/heaad/archive/2011/01/02/1924088.html)