特征选择方法综述

赖智豪

（江西中医药大学 计算机学院，江西 南昌330000）

摘要：特征选择在人工智能领域充当着重要角色，本文首先分析了特征选择的过程，将特征选择的过程分为一般过程与改进过程，两者进行对比分析。然后以搜索策略与评价准则划分特征选择的方法，基于搜索策略分为穷举搜索，启发式搜索和随机搜索，基于评价准则分为过滤法和封装法，并分别进行了分析与总结。

关键词：特征选择；搜索策略；评价准则

# 引 言

随着人工智能的发展，特征选择在机器学习，模式识别，数据挖掘等领域发挥着重要作用。特征选择( Feature Selection )也称特征子集选择( Feature Subset Selection , FSS )，或属性选择( Attribute Selection )。是指从已有的M个特征(Feature)中选择N个特征使得系统的特定指标最优化，是从原始特征中选择出一些最有效特征以降低数据集维度的过程,是提高学习算法性能的一个重要手段。[1] 简而言之，特征选择就是从已有的特征集合中选择一个最优特征子集的过程。

对于样本数据而言，有时候数据维度可能很高，高维度的数据特征空间虽然对事物的描述有一定的保障，但是难以避免当中含有冗余，不相关特征以及噪声特征。此时我们需要结合特征选择的方法在某个评价标准下选择出一个重要的特征子集。特征选择方法有很多，对于一个特定的问题，选择一个有效的特征选择方法极其重要。本文阐述了特征选择的过程，对比分析了特征选择的方法。

# 特征选择过程

## 一般的特征选择过程

特征选择过程一般包括这4部分：产生过程，评价函数，停止准则，验证过程。产生过程是搜索特征子集的过程，负责为评价函数提供特征子集。评价函数是评价一个特征子集的一个准则。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。验证过程是在验证数据集上检验特征子集有效性的过程。整个特征选择过程是：首先在原始数据中搜索，生成特征子集，然后根据评价函数对特征子集进行评价，评价结果如果符合停止准则就可以停止后续的搜索，然后进入结果检验阶段，否则返回重新搜索，生成特征子集，继续特征选择。整个特征选择过程如 图 1

|  |
| --- |
| **原始数据**  **生成特征子集**  **特征子集**  **子集评价**  **停止准则**  **结果检验**  **否**  **是** |
| 1. 一般的特征选择过程 |

## 改进的特征选择过程

由于特征子集搜索的时间复杂度较高。Yu等人[3] 基于相关冗余分析，得出了另一种改进的特征选择过程，它基本做法是：先对原始数据做相关分析，根据相关分析得出相关的特征子集，然后对相关特征子集作冗余分析，最终得出选择的特征子集。改进的特征选择过程如 图 2，改进的特征选择过程，它避免了特征子集搜索，经过相关分析和冗余分析可以高效快速的找出最优特征子集。

|  |
| --- |
| **原始数据**  **相关分析**  **特征子集**  **冗余分析**  **最终选择**  **的特征子集** |
| 1. 改进的特征选择过程 |

主要影响特征选择的是子集生成过程中的搜索策略和之后对子集的评价准则。所以本文就以搜索策略和评价准则划分特征选择的方法。

# 基于搜索策略的特征选择方法

在搜索特征子集过程中，基于搜索策略的特征选择方法又分为穷举搜索，启发式搜索，随机搜索。

### 穷举搜索

穷举搜索根据穷举集合是否完备可以划分为完备集穷举和非完备集穷举。完备集穷举，如广度优先法，遍历所有特征集合等。非完备集穷举，如分支界定法，最好优先法等。穷举搜索虽然能够将特征空间中的所有特征组合遍历到，但是它的计算复杂度很高。当有N个特征时，穷举搜索的计算复杂度达。在穷举搜索中，由于分支界定法中的剪枝操作可以减小计算复杂度，所以常用的是分支定界法。分支界定法是一种搜索和迭代的方法，它的基本思想是：在每次分支后，对凡是超出界限的已知可行解进行剪枝操作，那些子集不再做进一步分支。这样，解的许多子集（即搜索树上的许多结点）就可以不予考虑了，从而缩小了搜索范围。这一过程一直进行到找出可行解为止，该可行解的值不大于任何子集的界限。因此这种算法一般可以求得最优解。虽然分支界定法能找到最优解，但是在应用于高维度多分类问题中，它的效率还是很难达到要求。总之，穷举搜索由于诸多因素而无法广泛被应用。

### 启发式搜索

启发式搜索(Heuristically Search)又称为有信息搜索(Informed Search)，它是利用问题拥有的启发信息来引导搜索，达到减少搜索范围、降低问题复杂度的目的，这种利用启发信息的搜索过程称为启发式搜索。启发式搜索是一种近似算法,它可以有效解决组合爆炸问题，大大提高搜索效率。使用启发式搜索有两种基本的情况：一是问题陈述和数据获取方面具有模糊性，可能找不到这个问题的的确定解；二是，虽然问题可能有确定解，但是其状态空间特别大。

启发式搜索主要有：序列前向选择( SFS , Sequential Forward Selection )，序列后向选择( SBS , Sequential Backward Selection )，双向搜索( BDS , Bidirectional Search )，增L去R选择算法 ( LRS , Plus-L Minus-R Selection )，序列浮动选择( Sequential Floating Selection )，决策树( Decision Tree Method , DTM)。

序列前项选择(SFS)基本做法是：特征子集X由空集开始，每次选择一个能使评价函数J（X）最优的特征加入集合X。它是一种基于贪心的算法，因为它只加不减，所以它容易出现冗余。

序列后项选择(SBS)基本做法是：和SFS相反，特征集合从特征全集开始，每次剔除一个能使评价函数J(x)最优的特征。它也是一种基于贪心的算法，它只能减而不能加是它的缺点。SBS与SFS都容易陷入一个局部最优解。

双向搜索(BDS)基本做法是：结合SBS与SFS，从两端同时搜索，直到SBS与SFS搜索到同一个特征子集时停止。双向搜索相比于SBS和SFS，搜索速度有比较明显的提高，但是它并不能避免容易陷入局部最优解的情况。

增L去R选择算法基本做法有两种形式：一种是特征集从空集开始，每次选择L个特征，同时从L个中去掉R个特征，使得评价函数最优，将剩下的加入特征集。另一种是特征集从全集开始，每次去掉R个特征，从R个中选出L个特征，使评价函数最优，加入特征集。它也结合了SBS与SFS，该算法的关键是选择L与R。

序列浮动选择由增L去R发展而来，它的不同之处在于L与R不是固定的，而是“浮动”的。它又分为：序列浮动前向选择( SFFS, Sequential Floating Forward Selection )，序列浮动后向选择( SFBS, Sequential Floating Backward Selection )。序列浮动选择结合了序列前向选择、序列后向选择、增L去R选择的特点，并弥补了它们的缺点，它是一种比较常用的启发式搜索算法。

决策树基本做法：在训练样本集上运行C4.5或其他决策树生成算法，待决策树充分生长后，再在树上运行剪枝算法。最终决策树各分支处的特征就是选出来的特征子集。决策树方法一般使用信息增益作为评价函数。

### 随机搜索

在特征选择的随机搜索中，主要有：随机产生序列选择算法(RGSS, Random Generation plus Sequential Selection)，模拟退火算法( SA, Simulated Annealing )[5]，遗传算法( GA, Genetic Algorithms )[6]。随机搜索很可能跳出局部最优解，它们在一定程度上弥补了启发式搜索的缺点。但是，随机搜索太依赖于随机因素，有时难以求解。

在这些特征选择搜索方法中，每一种搜索方法都有各自的优缺点。在实际的应用中，应该根据具体情况而定。比如，当特征数量比较少时，可以考虑穷举搜索。当特征数量比较大，不需要全局最优且要计算速度快时，可以考虑用启发式搜索。当特征数量比较大，不需要考虑计算时间时可以使用随机搜索。

# 基于评价准则的特征选择方法

根据特征选择方法是否独立于后续的学习算法，将基于评价准则的特征选择方法分为过滤法（Filter）和封装法（Wrapper）。

#### 过滤法(Filter)

过滤法比较简单，它按照特征的发散性或者相关性指标对各个特征进行评分，设定评分阈值或者待选择阈值的个数，选择合适特征。过滤法可以独立于后续的学习算法，也就是说使用过滤法特征选择过程与学习器无关，相当于先对初始特征进行过滤，再用过滤后的特征训练学习器。这种方法的关键就是找到一种能度量特征重要性的方法。比如皮尔森相关系数，信息论理论中的互信息等。

常用的过滤式特征选择方法有：方差选择法，卡方检验法，皮尔森相关系数法，互信息系数法。

方差选择法：方差选择法适用于离散型特征，如果是连续型特征要作离散化处理。使用方差作为特征评分标准，如果某个特征取值差异不大，通常可以认为该特征对区分样本贡献不大。方差选择法是一种简单的方法，但是因为方差较小的特征比较少，所以它的使用范围不广泛。它一般用于特征选择预处理过程，先将原始数据中方差较小的特征去除而得到预处理数据，然后再使用其他特征选择方法选择特征。

卡方检验法：卡方检验法适用于分类问题。该方法使用统计量卡方检验作为特征选择评分标准，卡方检验值越大，相关性越强。

皮尔森相关系数法：皮尔森相关系数法能够衡量线性相关的特征集。使用pearson系数作为特征评分标准，相关系数绝对值越大，相关性越强（相关系数越接近于1或-1时，相关性越强；相关系数越接近于0时，相关性越弱）。

互信息系数法：互信息系数法也比较适用于分类问题。互信息系数法能够衡量各种相关性的特征集，但是它计算相对比较复杂。

对于过滤法特征选择，张平[7]基于相关性，提出加权相关性WR，同时提出了一种新颖的条件相关特征选择算法CRFS(Conditional Relevance Feature Selection),并在理论上证明了它的优越性，除此之外，从不同角度，在实验上取得了较好的结果。张尧[8]基于互信息提出了一种基于最小条件和最小条件冗余的特征选择算法(MCRMCR,Minimum Conditional Relevancy and Minimum Conditional Redundancy)和WCRFR(Weight Composition of Feature Relevancy)算法，并在理论与实际中证明这两个算法可以提高分类器的分类准确率。

#### 封装法(Wrapper)

封装法是根据目标函数，通常是预测效果评分，每次选择部分特征，或者排除部分特征。封装法的特征选择过程与学习器有关，使用学习器的性能作为评价准则，选择最有利于学习器性能的特征子集。

最常用的封装法是递归消除特征法(RFE，Recursive Feature Elimination)。递归消除特征法使用一个机器学习模型来进行多轮训练，每轮训练后，消除若干权值系数的对应的特征，再基于新的特征集进行下一轮训练。以经典的SVM-RFE算法来讨论这个特征选择的思路。这个算法以支持向量机来做RFE的机器学习模型选择特征。它在第一轮训练的时候，会选择所有的特征来训练，得到了分类的超平面后，如果有n个特征，那么RFE-SVM会选择出中分量的平方值最小的那个序号i对应的特征，将其排除，在第二类的时候，特征数就剩下n-1个了，我们继续用这n-1个特征和输出值来训练SVM，同样的，去掉最小的那个序号i对应的特征。以此类推，直到剩下的特征数满足我们的需求为止。对于该类算法的研究，王君[9]基于SVM-RFE，使用模拟退火结合皮尔森系数作为度量标准对已经删除的特征子集重新评估。林俊等人[11]提出了一种基于支持向量机递归特征消除法(SVM-RFE)和二进制粒子群算法(BPSO)的特征选择方法.该方法实验表明能够在分类准确率更高或相等的情况下得到维数更少的子集。封装法在速度上要比过滤法慢，但是它选择的特征子集的规模要小的多而且准确率比较高，有利于识别关键特征。它的缺点是时间复杂度较高，泛化能力不强。

综上所述，过滤法和封装法各有优缺点。但是过滤法和封装法可以是互补模式，两者结合起来能达到更好的效果。两者结合的特征选择分为两个阶段：第1阶段使用过滤法初步去除大部分无关或噪声特征，只保留少量特征，从而有效的减小后续搜索规模。第2阶段将剩余的特征连同样本数据作为输入参数用于封装法的相关操作，进一步优化选择重要特征。对于两者结合的算法研究，姜百宁[12]研究了两者相结合的特征选择方法，并基于互信息和遗传算法提出了MI-GA算法，该算法实验表明具有良好的综合性能。

# 结 论

本文结合文献首先阐述了特征选择的过程，然后从搜索策略和评价准则两个角度重点论述了特征选择的方法。特征选择的方法很多，但是需要结合实际情况进行使用。多种特征选择方法相结合的特征选择是一种值得的研究方向，比如对过滤法和封装法结合使用的研究。

**参考文献：**

1. 刘家锋，赵巍，朱海龙，金野编著；唐降龙主审 ．《模式识别》．哈尔滨：哈尔滨工业大学出版社，2014.
2. 姚旭，王晓丹，张玉玺，权文.特征选择方法综述，2012.
3. Lei Yu, Huan Liu. Efﬁcient feature selection via analysis of relevance and redundancy[J]. J of Machine Learning Research, 2004, 5(1): 1205-1224.
4. 王娟，慈林林，姚康泽.特征选择方法综述，2005.
5. 赵博研.基于改进模拟退火算法的项目选择优化方法研究,2019.
6. 李岩，袁弘宇，于佳乔，张更伟，刘克平.遗传算法在优化问题中的应用综述,2019.
7. 张平.基于过滤法的信息论特征选择算法研究,2018.
8. 张尧.基于互信息的特征选择方法研究,2019.
9. 王君.基于SVM-RFE的特征选择方法研究,2015.
10. Zhicheng Zhu; Zhiqiang Wei; Bo Yin; Tao Liu; Xianqing Huang.Feature selection of non-intrusive load monitoring system using RFE and RF,2018.
11. 林俊,许露,刘龙.基于SVM-RFE-BPSO算法的特征选择方法,2015.
12. 姜百宁.机器学习中的特征选择算法研究,2009.