特征选择算法及应用综述 *

刘飞飞
（山西大学商务学院 太原 030031）

摘 要 特征选择是根据某种策略和评价准则从原始数据特征集合中寻找最优特征子集的技术，它是针对高维复杂的数据进行预处理的重要方法。文章从特征选择的概念入手，介绍了特征选择算法的流程和分类方法，讨论了特征选择算法选用时需要考虑的因素，指出了特征选择算法的发展趋势及存在的问题。

关键词 特征选择 最优特征子集 预处理
中图分类号 TP393 **文献标识码** A **文章编号** 6550

A Survey of Feature Selection Algorithms and Applications

LIU Feifei
（Business College of Shanxi University Taiyuan 030031）

Abstract Feature selection is a technique to find the optimal feature subset from the original data feature set according to a certain strategy and evaluation criterion. It is an important method to preprocess high-dimensional complex data. Starting with the concept of feature selection, this paper introduces the process and classification method of feature selection algorithms, discusses the factors that need to be considered when selecting feature selection algorithms, and points out the development trend and existing problems of feature selection algorithms.

Keywords Feature selection Optimal feature subset Preprocess

一、引言

特征选择技术是数据降维的一种重要的方法，它的本质就是从原始数据最初的特征集合中，选取一组符合某种评定标准的最优的特征子集，以便在进行分类或回归等任务时，可以获得更好的模型，取得更加精确的分析结果。特征选择技术的研究最早出现在上个世纪 60 年代，主要用于解决与统计学及信号处理等相关的问题，数据中涉及到的特征数目较少，并且假定各个特征之间相互独立，通过对单个特征的评定来完成特征的选择，最终将特征进行组合来形成最优特征子集。由于没有考虑特征与分类以及特征与特征之间的相互关系，早期的特征选择算法在实际应用中的表现很不理想，后来出现的特征选择算法在这一方面做出了很大的改进。上个世纪 90 年代，针对大规模数据的机器学习的出现，使得传统的特征选择算法受到了严峻的考验，迫切需要准确性及运行效率高的新的算法；另外，特征选择技术还被广泛应用到与数据挖掘、模式识别等相关的实际问题中[1]。

二、特征选择的概念

同变量、属性等一样，特征也是数据的一个方面，它可以是离散型数据、连续型数据或布尔型数据等。在常见的分类问题中，特征可以被分为三类：相关特征，较大程度上影响分类的结果并且不能被取代的特征；无关特征，取值随机性强，对分类结果没有影响的特征；冗余特征，不会影响分类的结果或与其他的特征存在关联的特征。特征选择的任务就是要从输入的数据中，去除无用或冗余的特征，得到对分类最有价值的由相关特征组成的最优特征子集。

根据适用的实际的场合的不同，对于特征选择的定义也不尽相同。假设存在样本数据集 $T=\{S,F,C\}$，其中 $F=\{f_1,f_2,\cdots,f_n\}$ 表示数据的特征集合，$C=\{c_1,c_2,\cdots,c_n\}$ 表示数据的类别，$S=\{s_1,s_2,\cdots,s_n\}$ 则表示当前的数据样本集合。同时，用 $J(X):2^F\rightarrow[0,1]$ 来表示选择最优特征子集所使用的评价函数，它的值越大则代表选定的特征子集中所包含的信息量越多，此时，可以将常见的特征选择的定义归结为以下三种形式：

1.从原始特征集 F 中寻找使得 J(X)取值最大的特征子集；

2.设定一个阈值 J_t，在原始特征集 F 中寻找使得 $J(X)>J_t$，并且特征数目最少的特征子集；

* 基金项目：山西大学商务学院科研基金项目，项目名称入侵检测中特征提取技术的研究与应用，项目编号 2017013。

3.从原始特征集F中寻找使得J(X)的值尽可能的大，同时特征数目尽可能少的特征子集。

三、特征选择算法

1.算法的流程

根据1997年Dash和Liu提出的特征选择框架可知，一般的特征选择算法会包括子集生成、子集评价、停止条件和子集验证四个步骤[2]。子集生成是一个不断搜索的过程，以原始特征集合为基础，选择一个起点特征集合，采用特定的搜索策略按照一定的搜索方向，产生用于下一步评价的特征子集。子集评价则是通过某种评价准则对子集生成过程中产生的特征子集进行评价，判断其是否为最优特征子集，若是则替换当前的最优特征子集。停止条件是为了防止搜索过程进入无限循环而设置的，一般为搜索的次数或特征的数目的阈值。子集验证是特征选择算法的最后一个环节，通常会采用分类器分别对原始特征集合和选择得到的最优特征子集进行训练和测试，以比较验证选择得到的最优特征子集的优劣[3]。特征选择算法的伪代码如下所示：

输入：S：表示训练样本数据集，用n个特征组成的特征集Fs表示每一个样本数据s

Js：评价规则或函数

Gs：特征子集产生方法

输出：fs_{best}：选择得到的最优特征子集

Step1：初始化特征子集 fs=Start-items(Fs);// 起点特征集合

fsbest={根据Js从fs中选择得到最优特征子集};// 子集评价

Step2：DO BEGIN

fs=Search-strategy(fs,Gs,Fs);// 子集产生

fs'={根据Js从fs中选择得到最优特征子集};// 子集评价

IF(Js(fs')优于Js(fs_{best}))

fs_{best}=fs';

END UNTIL stop(Js,fs);// 终止条件

Step3：输出 fs_{best};

2.算法的分类

在特征选择的过程中，搜索策略及特征的评价十分关键，也经常将它们作为特征选择算法的分类依据。

(1)按照搜索策略进行分类

对于输入的n个原始特征集合来说，就存在2^n-1个不同的不为空的特征子集，这些特征子集构成了候选的特征搜索空间，从其中找寻最优特征子集所采用的搜索方法就是搜索策略。目前，根据特征选择算法所使用的搜索策略的不同，可以将其分为基于全局最优搜索策略、基于随机搜索策略和基于启发式搜索三种，如图1所示。

①基于全局最优搜索策略的特征选择算法。穷举法和分支定界法是全局最优式搜索主要采用的方法。穷举法也称为耗尽式搜索，它通过搜索每一个存在的特征子集，来发现并选取符合要求的最优的特征子集，例如回溯方法和变体方法等。由于它可以遍历所有的特征集合，所以一定可以找到全局范围内的最优的特征组合，但是当原始特征的数目较大时，搜索的空间将会很大，算法的执行效率也会随着降低，因此，实用性不强。分支定界法通过剪枝操作来缩短搜索所耗费的时间，也是目前为止全局最优搜索中唯一的可以获得最优结果的方法；但是，它要求在搜索开始前预先设定最优特征子集的数目，子集评价函数要满足单调性，同时，当等待处理的特征的维数较高时，要重复多次执行算法，这些都在很大程度上限制了它的应用。

②基于随机搜索策略的特征选择算法。它在搜索过程中，将特征选择与模拟退火、禁忌搜索、遗传算法等结合，以概率理论和采样过程为理论基础，基于分类的有效性对每一个候选特征进行权重的赋值，并根据用户定义的或自适应获取的阈值对特征的重要性进行评判，将权重超出阈值的特征选择做最优特征输出。随机搜索方法将分类性能用作特征评判的标准，具有较好的应用效果，但是，它的时间复杂度较高，不能保证算法输出的特征集合就是最优的特征子集。

③基于启发式搜索策略的特征选择算法。它是一种对搜索的最优性和计算量进行了折中考虑的近似算法，通过合理的启发规则的设计，重复迭代运算来产生最优的特征子集。根据起始特征集合和搜索方向的不同，可以分为单独最优特征选择、序列前向选择、序列后向选择和双向选择等。启发式搜索的复杂性低，执行效率高，在实际的应用中使用十分广泛；但是，在特征的选择搜索过程中，一旦某个特征被选择或者删除，将不能被撤回，这将容易导致局部最优解的产生。

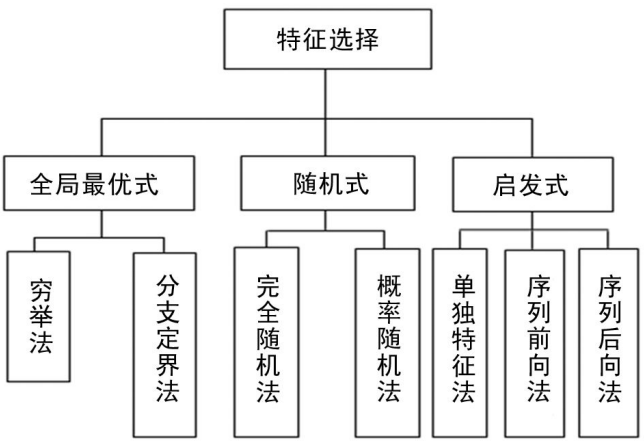

图1 特征选择的搜索策略

（2）按照评价策略进行分类

评价策略主要用来衡量特征选择算法性能或最优特征子集的优劣，根据所使用的评价策略可以将特征选择算法分为过滤式 Filter 和嵌入式 Wrapper 两种，二者的主要区别在于是否利用后续的分类器对特征子集进行评价：如果是，则为 Wrapper 方法，否则为 Filter 方法。

①基于 Filter 方法的特征选择算法。它是一种使用合适的评价准则来对特征的好坏进行评判的高效的方法，这些评价准则被用来增强特征和类之间的相关性或削弱不同特征之间的冗余性，目前，常用的评价准则有基于距离度量的、基于信息度量的、基于相关性度量的和基于一致性度量的等。这类特征选择方法的主要问题是无法保证能够找到的最优特征子集的规模是较小的，尤其是特征与分类器存在较大的关联性时，找到的满足条件的特征子集的规模会更大，存在的影响分类效果的噪声特征也更多。但是，由于基于 Filter 方法的特征选择方法可以依据相关的评价度量准则较快地剔除一部分噪声特征，减小特征搜索的空间，因此，可以用于对特征进行预选择处理。

②基于 Wrapper 方法的特征选择算法。这类算法对于特征子集的评价依赖于后续的分类算法，它直接使用已选择的特征子集来训练分类器，根据在测试集合上的分类效果来对特征子集的优劣进行判定，鉴于分类运算的复杂性，它的运行速度会比较慢，但是，它所选择的特征子集的规模比较小，分类的精度也比较高。分类错误率是基于 Wrapper 方法的特征选择算法性能优劣度量的主要标准。

鉴于基于 Filter 和 Wrapper 的特征选择方法各自的特点，在不同的应用场合，可以根据需要确定不同的评价准则，如表 1 所示，选用不同的特征选择算法，也可以将二者结合使用以获得更加精确的结果。

表 1 特征选择中不同的评价准则

评价准则	通用性	计算复杂性	分类准确率	适用的特征类型	代表算法
距离度量	强	低	不确定	连续、离散	Relief、ReliefF 等
信息度量	强	较高	不确定	连续、离散	MIFS、MRMR 等
相关性度量	强	低	不确定	连续、离散	MRMR 等
一致性度量	强	中等	不确定	离散	Focus、LVF 等
分类错误率度量	弱	很高	高	连续、离散	FFSR 等

四、特征选择算法的选用

目前，特征选择技术被广泛地应用于 Web 文档处理、图像处理、网络安全及医学诊断分析等领域，有关特征选择算法的研究也更加深入，大量新的算法不断涌现，算法的选用成为一个非常重要的问题[4]。通常来讲，特征选择算法的选用除了考虑应用的具体场景以外，还需要注意以下因素：第一，数据的规模。对于较小规模的数据集合，可以使用接近于完全搜索的 Filter 方法或 Wrapper 方法，如 BB 算法；而在数据集合的规模较大时，则应使用运行效率较高的 Filter 方法，如 Relief 或 ReliefF 算法等。第二，待处理的数据的类型。不同的特征选择算法可以处理的数据的类型也不尽相同，如 BB 算法不能处理离散型的数据，Relief 及 ReliefF 算法既可以处理离散型数据，也可以处理连续型数据，MIFS 及 MRMR 算法在处理连续型数据时，需要先对其进行离散化处理等。第三，待处理的数据的类别。对于类别未知的数据样本来说，应使用无监督的方法。第四，对分类器性能的要求。如果对分类器的输出精度的要求很高，则可以选用基于启发式搜索或基于遗传算法的 Wrapper 方法等。

五、小结

随着应用范围的不断拓展，针对特征选择算法的研究也越来越趋向于多样化和综合化，粗糙集、神经网络、支持向量机及模糊熵评价等不同领域的不同技术都被引入到特征选择中来；同时，非监督式的特征选择算法和将 filter 及 wrapper 融合的特征选择算法也成为研究和应用的热点。虽然，特征选择的相关理论及技术已经取得了很大的进展，但是，依然存在着很多有待研究和解决的难题，例如，比较有代表性的数据集的产生、特征选择与机器学习算法之间的关系的研究等。

参考文献

[1]李敏，卡米力.木依丁.特征选择方法与算法的研究[J].计算机技术与发展，2013，23（12）：16-21.

[2]计智伟，胡珉，尹建新.特征选择算法综述[J].电子设计工程，2011，19（9）：46-51.

[3]苏映雪.特征选择算法研究[D].长沙：国防科技大学，2006.

[4]张丽新，王家钦，赵雁南等.机器学习中的特征选择[J]，计算机科学，2004，31（11）：180-184.

作者简介

刘飞飞，女，1981 年 11 出生，河南洛阳人，计算机应用硕士研究生，讲师，研究方向为网络安全。