文章编号: 1001-0920 (2012) 02-0161-06

特征选择方法综述

姚 旭, 王晓丹, 张玉玺, 权 文

(空军工程大学 计算机工程系，陕西 三原 713800)

摘 要: 特征选择是模式识别的关键问题之一, 特征选择结果的好坏直接影响着分类器的分类精度和泛化性能. 首先分析了特征选择方法的框架; 然后从搜索策略和评价准则两个角度对特征选择方法进行了分析和总结; 最后分析了对特征选择的影响因素, 并指出了实际应用中需要解决的问题.

关键词: 特征选择；搜索策略；评价准则

中图分类号: TP391 **文献标识码:** A

Summary of feature selection algorithms

YAO Xu, *WANG Xiao-dan*, *ZHANG Yu-xi*, *QUAN Wen*

(Department of Computer Engineering，Air Force Engineering University，Sanyuan 713800，China. Correspondent: YAO Xu，E-mail：ffly132@163.com)

Abstract: Feature selection is one of the key processes in pattern recognition. The accuracy and generalization capability of classifier are affected by the result of feature selection directly. Firstly, the framework of feature selection algorithm is analyzed. Then feature selection algorithm is classified and analyzed from two points which are searching strategy and evaluation criterion. Finally, the problem is given to solve real-world applications by analyzing infection factors in the feature selection technology.

Key words: feature selection；searching strategy；evaluation criterion

1 引 言

特征选择是从一组特征中挑选出一些最有效的特征以降低特征空间维数的过程[1], 是模式识别的关键问题之一. 对于模式识别系统, 一个好的学习样本是训练分类器的关键, 样本中是否含有不相关或冗余信息直接影响着分类器的性能. 因此研究有效的特征选择方法至关重要.

本文分析讨论了目前常用的特征选择方法, 按照搜索策略和评价准则的不同对特征选择方法进行了分类和比较, 指出了目前特征选择方法及研究中存在的问题. 目前, 虽然特征选择方法有很多, 但针对实际问题的研究还存在很多不足, 如何针对特定问题给出有效的方法仍是一个需要进一步解决的问题.

2 特征选择的框架

迄今为止, 已有很多学者从不同角度对特征选择进行过定义: Kira 等人[2]定义理想情况下特征选择是寻找必要的、足以识别目标的最小特征子集; John 等人[3]从提高预测精度的角度定义特征选择是一个能够增加分类精度, 或者在不降低分类精度的条件下降低特征维数的过程; Koller 等人[4]从分布的角度定义特征选择为: 在保证结果类分布尽可能与原始数据类分布相似的条件下, 选择尽可能小的特征子集; Dash 等人[5]给出的定义是选择尽量小的特征子集, 并满足不显著降低分类精度和不显著改变类分布两个条件. 上述各种定义的出发点不同, 各有侧重点, 但是目标都是寻找一个能够有效识别目标的最小特征子集. 由文献 [2-5] 可知, 对特征选择的定义基本都是从分类正确率以及类分布的角度考虑. Dash 等人[5]给出了特征选择的基本框架, 如图 1 所示.

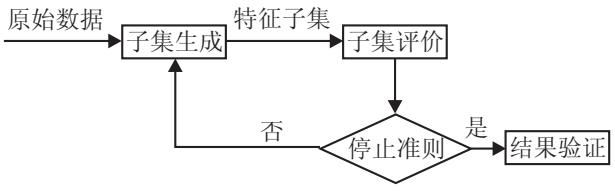

图 1 特征选择的基本框架

收稿日期: 2011-04-26；**修回日期:** 2011-07-12.

基金项目: 国家自然科学基金项目(60975026).

作者简介: 姚旭(1982−), 女, 博士生, 从事智能信息处理、机器学习等研究； 王晓丹(1966−), 女, 教授, 博士生导师, 从事智能信息处理、机器学习等研究.

由于子集搜索是一个比较费时的步骤, Yu 等人[6]基于相关和冗余分析, 给出了另一种特征选择框架, 避免了子集搜索, 可以高效快速地寻找最优子集. 框架如图 2 所示.

图 2 改进的特征选择框架

从特征选择的基本框架可以看出, 特征选择方法中有 4 个基本步骤: 候选特征子集的生成 (搜索策略)、评价准则、停止准则和验证方法[7-8]. 目前对特征选择方法的研究主要集中于搜索策略和评价准则, 因而, 本文从搜索策略和评价准则两个角度对特征选择方法进行分类.

3 基于搜索策略划分特征选择方法

基本的搜索策略按照特征子集的形成过程可分为以下 3 种: 全局最优、随机搜索和启发式搜索[9]. 一个具体的搜索算法会采用两种或多种基本搜索策略, 例如遗传算法是一种随机搜索算法, 同时也是一种启发式搜索算法. 下面对 3 种基本的搜索策略进行分析比较.

3.1 采用全局最优搜索策略的特征选择方法

迄今为止, 唯一得到最优结果的搜索方法是分支定界法[10]. 这种算法能保证在事先确定优化特征子集中特征数目的情况下, 找到相对于所设计的可分性判据而言的最优子集. 它的搜索空间是 $O(2^N)$ (其中 N 为特征的维数). 存在的问题: 很难确定优化特征子集的数目; 满足单调性的可分性判据难以设计; 处理高维多类问题时, 算法的时间复杂度较高. 因此, 虽然全局最优搜索策略能得到最优解, 但因为诸多因素限制, 无法被广泛应用.

3.2 采用随机搜索策略的特征选择方法

在计算过程中把特征选择问题与模拟退火算法、禁忌搜索算法、遗传算法等, 或者仅仅是一个随机重采样[11-12]过程结合起来, 以概率推理和采样过程作为算法的基础, 基于对分类估计的有效性, 在算法运行中对每个特征赋予一定的权重; 然后根据用户所定义的或自适应的阈值来对特征重要性进行评价. 当特征所对应的权重超出了这个阈值, 它便被选中作为重要的特征来训练分类器. Relief 系列算法即是一种典型的根据权重选择特征的随机搜索方法, 它能有效地去掉无关特征, 但不能去除冗余, 而且只能用于两类分类. 随机方法可以细分为完全随机方法和概率随机方法两种. 虽然搜索空间仍为 $O(2^N)$, 但是可以通过设置最大迭代次数限制搜索空间小于 $O(2^N)$. 例如遗传算法, 由于采用了启发式搜索策略, 它的搜索空间远远小于 $O(2^N)$.

存在的问题: 具有较高的不确定性, 只有当总循环次数较大时, 才可能找到较好的结果. 在随机搜索策略中, 可能需对一些参数进行设置, 参数选择的合适与否对最终结果的好坏起着很大的作用. 因此, 参数选择是一个关键步骤.

3.3 采用启发式搜索策略的特征选择方法

这类特征选择方法主要有: 单独最优特征组合, 序列前向选择方法 (SFS), 广义序列前向选择方法 (GSFS), 序列后向选择方法 (SBS), 广义序列后向选择方法 (GSBS), 增 l 去 r 选择方法, 广义增 l 去 r 选择方法, 浮动搜索方法. 这类方法易于实现且快速, 它的搜索空间是 $O(N^2)$. 一般认为采用浮动广义后向选择方法 (FGSBS) 是较为有利于实际应用的一种特征选择搜索策略, 它既考虑到特征之间的统计相关性, 又用浮动方法保证算法运行的快速稳定性[13]. 存在的问题是: 启发式搜索策略虽然效率高, 但是它以牺牲全局最优为代价.

每种搜索策略都有各自的优缺点, 在实际应用过程中, 可以根据具体环境和准则函数来寻找一个最佳的平衡点. 例如, 如果特征数较少, 可采用全局最优搜索策略; 若不要求全局最优, 但要求计算速度快, 则可采用启发式策略; 若需要高性能的子集, 而不介意计算时间, 则可采用随机搜索策略.

4 基于评价准则划分特征选择方法

特征选择方法依据是否独立于后续的学习算法, 可分为过滤式 (Filter) 和封装式 (Wrapper)[14]两种. Filter 与后续学习算法无关, 一般直接利用所有训练数据的统计性能评估特征, 速度快, 但评估与后续学习算法的性能偏差较大. Wrapper 利用后续学习算法的训练准确率评估特征子集, 偏差小, 计算量大, 不适合大数据集. 下面分别对 Filter 和 Wrapper 方法进行分析.

4.1 过滤式 (Filter) 评价策略的特征选择方法

Filter 特征选择方法一般使用评价准则来增强特征与类的相关性, 削减特征之间的相关性. 可将评价函数分成 4 类[5]: 距离度量、信息度量、依赖性度量以及一致性度量.

4.1.1 距离度量

距离度量通常也认为是分离性、差异性或者辨识能力的度量. 最为常用的一些重要距离测度[1]有欧氏距离、S 阶 Minkowski 测度、Chebychev 距离、平方距离等. 两类分类问题中, 对于特征 X 和 Y, 如果由 X 引起的两类条件概率差异性大于 Y, 则 X 优于 Y. 因为特征选择的目的是找到使两类尽可能分离的

特征. 如果差异性为 0, 则 X 与 Y 是不可区分的. 算法 Relief[2] 及其变种 ReliefF[15], 分支定界 和 BFF[16]等都是基于距离度量的. 准则函数要求满足单调性, 也可通过引进近似单调的概念放松单调性的标准. 蔡哲元等人[17]提出了基于核空间的距离度量, 有效地提高了小样本与线性不可分数据集上的特征选择能力.

4.1.2 信息度量

信息度量通常采用信息增益 (IG) 或互信息 (MI) 衡量. 信息增益定义为先验不确定性与期望的后验不确定性之间的差异, 它能有效地选出关键特征, 剔除无关特征[18]. 互信息描述的是两个随机变量之间相互依存关系的强弱. 信息度量函数 $J(f)$ 在 Filter 特征选择方法中起着重要的作用. 尽管 $J(f)$ 有多种不同形式, 但是目的是相同的, 即使得所选择的特征子集与类别的相关性最大, 子集中特征之间的相关性最小. 刘华文[19]给出了一种泛化的信息标准, 准则如下:

$$J(f) = \alpha \cdot g(C, f, S) - \delta. \tag{1}$$

其中: C 为类别, f 为候选特征, S 为已选择的特征, 函数 $g(C, f, S)$ 为 C, f, S 之间的信息量; α 为调控系数, δ 为惩罚因子. 下面就此信息标准的泛化形式与几个现有选择算法中的信息度量标准之间的关系进行讨论:

1) BIF (best individual feature)[20]是一种最简单最直接的特征选择方法. 它的评价函数为

$$J(f) = I(C; f), \tag{2}$$

其中 $I(\,)$ 为互信息, $I(C; f)$ 为类别 C 与候选特征 f 之间的互信息. 它的基本思想是对于每一个候选特征 f 计算评价函数 $J(f)$, 并按评价函数值降序排列, 取前 k 个作为所选择的特征子集. 这种方法简单快速, 尤其适合于高维数据. 但是它没有考虑到所选特征间的相关性, 会带来较大的冗余.

2) MIFS (mutual information feature selection) 为基于上述算法的缺点, 由 Battiti[21]给出的一种使用候选特征 f 与单个已选特征 s 相关性对 f 进行惩罚的方法, 其评价函数为

$$J(f) = I(C; f) - \beta \sum_{s \in S} I(s; f), \tag{3}$$

其中 β 为调节系数, 当 $\beta \in [0.5, 1]$ 时, 算法性能较好.

3) mRMR (minimal-redundancy and maximal-relevance) [22]方法. 从理论上分析了 mRMR 等价于最大依赖性, 并分析了三者的关系. 基于最大依赖性, 可通过计算不同特征子集与类别的互信息来选取最优子集. 但是, 在高维空间中, 估计多维概率密度是一个难点. 另一个缺点是计算速度非常慢. 所以本文从与其等价的最小冗余和最大相关出发, 给出一种基于互信息的评价准则, 具体函数如下:

$$J(f) = I(C; f) - \frac{1}{|S|} \sum_{s \in S} I(s; f), \tag{4}$$

其中 $|S|$ 表示已选特征的个数. 该算法的思想就是最大化特征子集和类别的相关性, 同时最小化特征之间的冗余. Peng 用这种方法将多变量联合概率密度估计问题转化为多重二变量概率密度估计, 解决了一大难题. Ding等人[23]还给出了此算法的一种变种形式, 将准则函数中的减法改为除法, 即

$$J(f) = \frac{I(C; f)}{\dfrac{1}{|S|} \displaystyle\sum_{s \in S} I(s; f)}. \tag{5}$$

4) FCBF (fast correlation-based filter)[6]是基于相互关系度量给出的一种算法. 对于线性随机变量, 用相关系数分析特征与类别、特征间的相互关系. 对于非线性随机变量, 采用对称不确定性 (SU) 来度量. 对于两个非线性随机变量 X 和 Y, 它们的相互关系可表示为

$$SU(X, Y) = 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right]. \tag{6}$$

其中: $H(X)$ 与 $H(Y)$ 为信息熵, $IG(X|Y)$ 为信息增益. 该算法的基本思想是根据所定义的 C- 相关 (特征与类别的相互关系) 和 F- 相关 (特征之间的相互关系), 从原始特征集合中去除 C- 相关值小于给定阈值的特征, 再对剩余的特征进行冗余分析.

5) CMIM (conditional mutual information maximization). 有些特征选择方法利用条件互信息来评价特征的重要性程度, 即在已知已选特征集 S 的情况下通过候选特征 f 与类别 C 的依赖程度来确定 f 的重要性, 其中条件互信息 $I(C; f|S)$ 值越大, f 能提供的新信息越多. 因为 $I(C; f|S)$ 计算费用较高, 且样本的多维性导致了其估值不准确, Fleuret[24]在提出的条件互信息最大化选择算法 CMIM 中采取一种变通的方式, 即使用单个已选特征 s 来代替整个已选子集 S 以估算 $I(C; f|S)$, 其中 s 是使 $I(C; f|S)$ 值最大的已选特征. CMIM 的评价函数为

$$J(f) = \arg\min_{s \in S} I(C; f|S). \tag{7}$$

除以上几种信息度量和算法外, 针对存在的问题, 研究者们提出了新的评价函数和算法. Kwak 等人[25]指出 MIFS 算法中评价函数 $J(f)$ 的惩罚因子并不能准确地表达冗余程度的增长量, 给出了 MIFS-U (MIFS-uncertainty) 算法; 与 MIFS 算法类似, MIFS-U 算法中参数 β 的取值将影响选择算法的性能. 为了解决该问题, Novovicova 等人[26]提出了 MIFS-U 的一种改进算法 mMIFS-U (modified version of MIFS-U), 算法中将 f 与 S 中单个已选特征相关程度最大的 s 作为它们之间的冗余程度; 为了解决对称不确定性可能

提供一些错误或不确定信息, Qu 等人[27]利用决策依赖相关性来精确度量特征 f 与 s 间的依赖程度, 提出了 DDC (decision dependent correlation) 算法. 它们的思想都是一致的, 只是评价函数的表达形式不同. 刘华文[19]还提出了一种基于动态互信息的特征选择方法. 随着已选特征数的增加, 类别的不确定性也逐渐降低, 无法识别的样本数也越来越少. 因此, 已识别的样本会给特征带来干扰信息, 可采用动态互信息作为评价标准, 在特征选择过程中不断地删除已识别的样本, 使得评价标准在未识别样本上动态估值.

基于信息的度量是近年来的一个研究热点, 出现了大量基于信息熵的特征选择方法, 如文献 [28-31] 等. 因为信息熵理论不要求假定数据分布是已知的, 能够以量化的形式度量特征间的不确定程度, 并且能有效地度量特征间的非线性关系. 因此, 信息度量被广泛应用, 并且也通过试验证明了其性能[32-34]. 以上基于信息度量的评价准则虽然形式不同, 但是核心思想都是找出与类别相关性最大的特征子集, 并且该子集中特征之间的相关性最小. 设计体现这一思想的函数是至关重要的.

4.1.3 依赖性度量

有许多统计相关系数, 如 Pearson 相关系数、概率误差、Fisher 分数、线性可判定分析、最小平方回归误差[35]、平方关联系数[36]、t-test 和 F-Statistic 等被用来表达特征相对于类别可分离性间的重要性程度. 例如, Ding[23]和 Peng[22]在 mRMR 中处理连续特征时, 分别使用 F-Statistic 和 Pearson 相关系数度量特征与类别和已选特征间的相关性程度. Hall[37]给出一种既考虑了特征的类区分能力, 同时又考虑特征间冗余性的相关性度量标准. Zhang 等人[38]使用成对约束即 must-link 约束和 cannot-link 约束计算特征的权重, 其中 must-link 约束表示两个样本离得很近, 而 cannot-link 表示样本间离得足够远.

在依赖性度量中, Hilbert-Schmidt 依赖性准则 (HSIC) 可作为一个评价准则度量特征与类别的相关性. 核心思想是一个好的特征应该最大化这个相关性. 特征选择问题可以看成组合最优化问题

$$T_0 = \arg\max_{S\subseteq F} J(S),\ \text{s.t.}|S| \leqslant t. \tag{8}$$

其中: t 为所选特征个数的上限, F 为特征集合, S 为已选特征的集合, $J(S)$ 为评价准则. 从式 (8) 中可知需要解决两个问题: 一是评价准则 $J(S)$ 的选择; 二是算法的选择. 文献 [39-40] 是 HSIC 准则的具体应用.

4.1.4 一致性度量

给定两个样本, 若他们特征值均相同, 但所属类别不同, 则称它们是不一致的; 否则, 是一致的. 一致性准则用不一致率来度量, 它不是最大化类的可分离性, 而是试图保留原始特征的辨识能力, 即找到与全集有同样区分类别能力的最小子集. 它具有单调、快速、去除冗余和不相关特征、处理噪声等优点, 能获得一个较小的特征子集. 但其对噪声数据敏感, 且只适合离散特征. 典型算法有 Focus[41], LVF[42]等. 文献 [43-44] 给出了基于不一致度量的算法.

上面分析了 Filter 方法中的一些准则函数, 选择合适的准则函数将会得到较好的分类结果. 但 Filter 方法也存在很多问题: 它并不能保证选择出一个优化特征子集, 尤其是当特征和分类器息息相关时. 因而, 即使能找到一个满足条件的优化子集, 它的规模也会比较庞大, 会包含一些明显的噪声特征. 但是它的一个明显优势在于可以很快地排除很大数量的非关键性的噪声特征, 缩小优化特征子集搜索的规模, 计算效率高, 通用性好, 可用作特征的预筛选器.

4.2 封装式 (Wrapper) 评价策略的特征选择方法

除了上述 4 种准则, 分类错误率也是一种衡量所选特征子集优劣的度量标准. Wrapper 模型将特征选择算法作为学习算法的一个组成部分, 并且直接使用分类性能作为特征重要性程度的评价标准. 它的依据是选择子集最终被用于构造分类模型. 因此, 若在构造分类模型时, 直接采用那些能取得较高分类性能的特征即可, 从而获得一个分类性能较高的分类模型. 该方法在速度上要比 Filter 方法慢, 但是它所选择的优化特征子集的规模相对要小得多, 非常有利于关键特征的辨识; 同时它的准确率比较高, 但泛化能力比较差, 时间复杂度较高. 目前此类方法是特征选择研究领域的热点, 相关文献也很多. 例如, Hsu 等人[45]用决策树来进行特征选择, 采用遗传算法来寻找使得决策树分类错误率最小的一组特征子集. Chiang 等人[46]将 Fisher 判别分析与遗传算法相结合, 用来在化工故障过程中辨识关键变量, 取得了不错的效果. Guyon 等人[47]使用支持向量机的分类性能衡量特征的重要性程度, 并最终构造一个分类性能较高的分类器. Krzysztof [48]提出了一种基于相互关系的双重策略的 Wrapper 特征选择方法. 叶吉祥等人[49]提出了一种快速的 Wrapper 特征选择方法 FFSR(fast feature subset ranking), 以特征子集作为评价单位, 以子集收敛能力作为评价标准. 戴平等人[50]利用 SVM 线性核与多项式核函数的特性, 结合二进制 PSO 方法, 提出了一种基于 SVM 的快速特征选择方法.

综上所述, Filter 和 Wrapper 特征选择方法各有优缺点. 将启发式搜索策略和分类器性能评价准则相结合来评价所选的特征, 相对于使用随机搜索策略的方法, 节约了不少时间. Filter 和 Wrapper 是两种

互补的模式, 两者可以结合. 混合特征选择过程一般由两个阶段组成, 首先使用Filter方法初步剔除大部分无关或噪声特征, 只保留少量特征, 从而有效地减小后续搜索过程的规模. 第2阶段将剩余的特征连同样本数据作为输入参数传递给Wrapper选择方法, 以进一步优化选择重要的特征. 例如, 文献[51]采用混合模型选择特征子集, 先使用互信息度量标准和bootstrap技术获取前k个重要的特征, 然后再使用支持向量机构造分类器.

5 结　论

本文首先分析了特征选择的框架, 然后从两个角度对特征选择方法进行分类: 一个是搜索策略, 一个是评价准则. 特征选择方法从研究之初到现在, 已经有了很多成熟的方法, 但是, 研究过程中也存在很多问题. 例如: 如何解决高维特征选择问题; 如何设计小样本问题的特征选择方法; 如何针对不同问题设计特定的特征选择方法; 研究针对新数据类型的特征选择方法等. 影响特征选择方法的因素主要有数据类型、样本数量. 针对两类还是多类问题, 特征选择方法的选择也有不同. 例如Koll-Saha[4]和Focus等人[41]受限于连续型特征; 分支定界, BFF[16]和MDLM(min description length method)[52]等不支持布尔型特征; Relief系列算法, DTM(decision tree method)[53]和PRESET[54]都适合于大数据集; Focus等人[41]适用于小样本; 在度量标准的选择中, 只有一致性度量仅适用于离散型数据等等.

尽管特征选择方法已经有很多, 但针对解决实际问题的方法还存在很多不足, 如何针对特定问题给出有效的方法仍是一个需要进一步解决的问题. 将Filter方法和Wrapper方法两者结合, 根据特定的环境选择所需要的度量准则和分类器是一个值得研究的方向.

参考文献(References)

[1] 边肇祺, 张学工. 模式识别[M]. 第2版. 北京: 清华大学出版社, 2000.
(Bian Z Q, Zhang X G. Pattern recognition[M]. 2nd ed. Beijing: Tsinghua University Publisher, 2000.)

[2] Kira K, Rendell L A . The feature selection problem: Traditional methods and a new algorithm[C]. Proc of the 9th National Conf on Artificial Intelligence. Menlo Park, 1992: 129-134.

[3] John G H, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem[C]. Proc of the 11th Int Conf on Machine Learning. New Brunswick, 1994: 121-129.

[4] Koller D, Sahami M. Toward optimal feature selection[C]. Proc of Int Conf on Machine Learning. Bari, 1996: 284-292.

[5] Manoranjan Dash, Huan Liu. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(3): 131-156.

[6] Lei Yu, Huan Liu. Efficient feature selection via analysis of relevance and redundancy[J]. J of Machine Learning Research, 2004, 5(1): 1205-1224.

[7] Liu H, Motoda H. Feature selection for knowledge discovery and data mining[M]. Boston: Kluwer Academic Publishers, 1998.

[8] Molina L C, Lluís Belanche, Àngela Nebot. Feature selection algorithms: A survey and experimental evaluation[R]. Barcelona: Universitat Politecnicade Catalunya, 2002.

[9] Sun Z H, George Bebis, Ronald Miller. Object detection using feature subset selection[J]. Pattern Recognition, 2004, 37(11): 2165-2176.

[10] Narendra P M, Fukunaga K. A branch and bound algorithm for feature selection[J]. IEEE Trans on Computers, 1977, 26(9): 917-922.

[11] Tsymbal A, Seppo P, David W P. Ensemble feature selection with the simple Bayesian classification[J]. Information Fusion, 2003, 4(2): 87-100.

[12] Wu B L, Tom A, David F, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data[J]. Bioinformatics, 2003, 19(13): 1636-1643.

[13] Furlanello C, Serafini M, Merler S, et al. An accelerated procedure for recursive feature ranking on microarray data[J]. Neural Networks, 2003, 16(4): 641-648.

[14] Langley P. Selection of relevant features in machine learning[C]. Proc of the AAAI Fall Symposium on Relevance. New Orleans, 1994: 1-5.

[15] Kononenko I. Estimation attributes: Analysis and extensions of RELIEF[C]. Proc of the 1994 European Conf on Machine Learning. New Brunswick, 1994: 171-182.

[16] Xu L, Yan P, Chang T. Best first strategy for feature selection[C]. Proc of 9th Int Conf on Pattern Recognition. Rome, 1988: 706-708.

[17] 蔡哲元, 余建国, 李先鹏, 等. 基于核空间距离测度的特征选择[J]. 模式识别与人工智能, 2010, 23(2): 235-240.
(Cai Z Y, Yu J G, Li X P, et al. Feature selection algorithm based on kernel distance measure[J]. Pattern Recognition and Artificial Intelligence, 2010, 23(2): 235-240.)

[18] 徐燕, 李锦涛, 王斌, 等. 基于区分类别能力的高性能特征选择方法[J]. 软件学报, 2008, 19(1): 82-89.
(Xu Y, Li J T, Wang B, et al. A category resolve power-based feature selection method[J]. J of Software, 2008, 19(1): 82-89.)

[19] 刘华文. 基于信息熵的特征选择算法研究[D]. 长春: 吉林大学, 2010.
(Liu Hua-wen. A study on feature selection algorithm using information entropy[D]. Changchun: Jilin University, 2010.)

[20] Jain A K, Robert P W, Mao J C. Statistical pattern recognition: A review[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.

[21] Battiti R. Using mutual information for selecting features in supervised neural net learning[J]. IEEE Trans on Neural Networks, 1994, 5(4): 537-550.

[22] Hanchuan Peng, Fuhui Long, Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[23] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. J of Bioinformatics and Computational Biology, 2005, 3(2): 185-205.

[24] Francois Fleuret. Fast binary feature selection with conditional mutual information[J]. J of Machine Learning Research, 2004, 5(10): 1531-1555.

[25] Kwak N, Choi C-H. Input feature selection by mutual information based on Parzen window[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(12): 1667-1671.

[26] Novovicova J, Petr S, Michal H, et al. Conditional mutual information based feature selection for classification task[C]. Proc of the 12th Iberoamericann Congress on Pattern Recognition. Valparaiso, 2007: 417-426.

[27] Qu G, Hariri S, Yousif M. A new dependency and correlation analysis for features[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(9): 1199-1207.

[28] 赵军阳, 张志利. 基于模糊粗糙集信息熵的蚁群特征选择方法[J]. 计算机应用, 2009, 29(1): 109-111.
(Zhao J Y, Zhang Z L. Ant colony feature selection based on fuzzy rough set information entropy[J]. J of Computer Applications, 2009, 29(1): 109-111.)

[29] 赵军阳, 张志利. 基于最大互信息最大相关熵的特征选择方法[J]. 计算机应用研究, 2009, 26(1): 233-235.
(Zhao J Y, Zhang Z L. Feature subset selection based on maxmutual information and max correlation entropy[J]. Application Research of Computers, 2009, 26(1): 233-235.)

[30] 渠小洁. 一种基于条件熵的特征选择算法[J]. 太原科技大学学报, 2010, 31(5): 413-416.
(Qu X J. An algorithm of feature selection based on conditional entropy[J]. J of Taiyuan University of Science and Technology, 2010, 31(5): 413-416.)

[31] 孟洋, 赵方. 基于信息熵理论的动态规划特征选取算法[J]. 计算机工程与设计, 2010, 31(17): 3879-3881.
(Meng Y, Zhao F. Feature selection algorithm based on dynamic programming and comentropy[J]. Computer Engineering and Design, 2010, 31(17): 3879-3881.)

[32] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. J of Machine Learning Research, 2003, 3(11): 1289-1305.

[33] Liu H, Liu L, Zhang H. Feature selection using mutual information: An experimental study[C]. Proc of the 10th Pacific Rim Int Conf on Artificial Intelligence. Las Vegas, 2008: 235-246.

[34] Hua J, Waibhav D T, Edward R D. Performance of feature-selection methods in the classification of high-dimension data[J]. Pattern Recognition, 2009, 42(7): 409-424.

[35] Mitra P, Murthy C A, Sankar K P. Unsupervised feature selection using feature similarity[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.

[36] Wei H-L, Billings S A. Feature subset selection and ranking for data dimensionality reduction[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(1): 162-166.

[37] Hall M A. Correlation-based feature subset selection for machine learning[M]. Hamilton: University of Waikato, 1999.

[38] Zhang D, Chen S, Zhou Z-H. Constraint score: A new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41(5): 1440-1451.

[39] Le Song, Alex Smola, Arthur Gretton, et al. Supervised feature selection via dependence estimation[C]. Proc of the 24th Int Conf on Machine Learning. Corvallis, 2007: 245-252.

[40] Gustavo Camps-Valls, Joris Mooij, Bernhard Scholkopf. Remote sensing feature selection by kernel dependence measures[J]. IEEE Geoscience and Remote Sensing Letters, 2010, 7(3): 587-591.

[41] Almuallim H, Dietterich T G. Learning with many irrelevant features[C]. Proc of 9th National Conf on Artificial Intelligence. Menlo Park, 1992: 547-552.

[42] Liu H, Setiono R. A probabilistic approach to feature selection–A filter solution[C]. Proc of Int Conf on Machine Learning. Bari, 1996: 319-327.

[43] Manoranjan Dash, Huan Liu. Consistency-based search in feature selection[J]. Artificial Intelligence, 2003, 151(16): 155-176.

[44] Huan Liu, Hiroshi Motoda, Manoranjan Dash. A monotonic measure for optimal feature selection[M]. Machine Learning: ECML-98, Lecture Notes in Computer Science, 1998: 101-106.

(下转第192页)