特征选择算法研究综述*

毛 勇[1] 周晓波[2] 夏 铮[1] 尹 征[1] 孙优贤[1]

[1](浙江大学 信息科学与工程学院 杭州 310027)
[2](哈佛大学 医学院 波士顿 美国 02115)

摘 要 特征选择是当前信息领域,尤其是模式识别领域的研究热点.本文从不同角度对特征选择算法进行分类,概述特征选择技术发展的各个分支及发展态势,指出理论研究和实际应用中所存在的一些困难和亟待解决的问题.然后从算法实用性角度出发,结合机器学习的观点,探讨应用支持向量机技术进行特征选择的研究发展思路.

关键词 特征选择,模式识别,机器学习,Wrapper 方法
中图法分类号 O235

A Survey for Study of Feature Selection Algorithms

MAO Yong[1], ZHOU Xiao-Bo[2], XIA Zheng[1], Yin Zheng[1], SUN You-Xian[1]
[1](*College of Information Science and Engineering, Zhejiang University, Hangzhou* 310027)
[2](*Medical School, Harvard University, Boston* 02115, *USA*)

ABSTRACT

Feature selection is a hot topic in current information science , especially in the field of pattern recognition. In this paper, feature selection algorithms are classified from different points of view. Several embranchments of feature selection and the development situation are introduced. Some difficulties in the theoretic analysis and application are involved. From a practicality angle, using support vector machine to select features is considered as the research direction in machine learning.

Key Words Feature Selection, Pattern Recognition, Machine Learning, Wrapper Method

* 国家 973 计划项目(No. 2002CB312200)、国家自然科学基金项目(No. 60574019、No. 60474045)、浙江省科技计划重点项目(No. 2005C21087)和浙江省科技计划院士基金项目(No. 2005A1001-13)资助
收稿日期:2006-02-24;修回日期:2006-07-24
作者简介 **毛勇**,男,1979 年生,博士,主要研究方向为模式识别与人工智能技术在医药信息学与生物信息学领域的应用. E-mail: ymao@iipc.zju.edu.cn. **周晓波**,男,1966 年生,博士,主要研究方向为生物信息学与信号处理方法在面向药物开发与诊疗的高分辨率分子与细胞图像处理领域的应用. **夏铮**,男,1980 年生,博士研究生,主要研究方向为模式识别与人工智能技术在高分辨率分子与细胞图像处理的应用. **尹征**,男,1981 年生,博士研究生,主要研究方向为统计模式识别方法在生物信息学领域的应用. **孙优贤**,男,1940 年生,教授,博士生导师,院士,主要研究方向为控制理论及应用、复杂系统建模、控制与优化.

1 引 言

特征选择是模式识别领域的研究热点之一. 模式识别系统的设计组成大致可以分为模式获取,预处理,特征选择/提取,回归/分类/描述,后处理这 5 个模块. 它的流程可以用图 1 来进行表述[1]. 特征选择在模式识别当中占有重要地位.

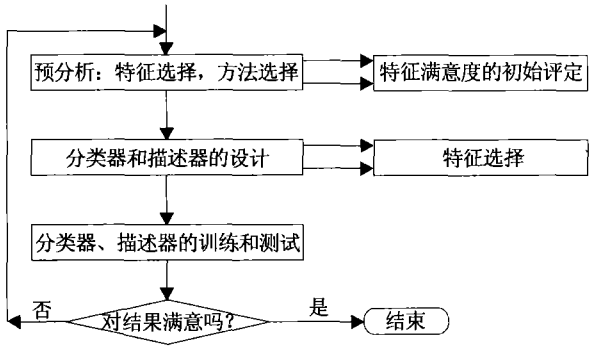

图 1 模式识别系统设计组成基本流程图

Fig. 1 Basic flowchart for design and construction of pattern recognition systems

从分类的角度看,模式识别是把具体事物归到具体某一类别的过程. 也就是先用一定数量的样本,根据它们之间的相似性进行分类器设计,而后用所设计的分类器对待识别的样本进行分类决策. 分类过程既可以在原始数据空间中进行,也可以对原始数据进行变换,将数据映像到最能反映分类本质的特征空间中进行. 相比而言,后者使得决策机器的设计更为容易,它通过更为稳定的特征表示,提高决策机器的性能,删去多余或不相关的信息,并且更加容易发现研究对象之间的固有联系. 因而,特征是决定样本之间的相似性和分类器设计的关键[2]. 在分类目的决定之后,如何找到合适的特征是认知与识别的核心问题. 但是,由于在很多实际问题当中,常常不容易找到那些最重要的特征,或者受条件限制不能对它们进行测量,这使得特征选择和提取的任务复杂化,从而成为构造模式识别系统、提高决策精度的最困难的任务之一[3-4].

模式识别的另一项关键任务是进行数据分析和处理. 一般来说,模式识别系统中包含由特征和属性所描述的对象的数学模型,研究一般意义上的对象间的相互关系很大程度上依赖于特征. 在认知科学中,很多状况下对象间的关系是已知的,而导致这种对象间显著差别的原因是未知的. 一般状态下,可能是几个关键性的特征在引导着对象间的差别[5-7]. 例如,在实际应用的过程中,不同对象可能代表着正常状况、疾病或故障状况,靠这些特征就可以表述正常状态、故障或疾病状态之间的差别. 如果能够甄别出这些导致区别的关键特征,就可以直接或间接找到解决对象间差异的办法,例如找到导致故障或者疾病的关键原因,可以为系统恢复正常提供指导[8]. 由于上述原因,特征选择在很多领域已经越来越受到人们的重视.

2 特征选择方法的分类及特性

特征选择的任务是从一组数量为 D 的特征中选择出数量为 $d(D>d)$ 的一组最优特征. 由于各个特征之间存在复杂的相互关系,在大多数情况下,如果仅对每个单独的特征按照一定的统计或者可分性判据进行排队,取排在前面的 d 个特征,所取得的结果在大多数情况下不是最优特征组,在仿真状况下甚至还有可能取到最差的特征组[3,9]. 从 D 个特征中选择出 d 个最优的特征,在这两个参数都已知的状况下,所有可能的组合数为

$$Q = C_D^d = D!/[(D-d)! \cdot d!].$$

如果 $D=100, d=10$,则 Q 的数量级是 10^{13},若 $D=20, d=10$,则 $Q=184756$. 如果把各种可能的特征组合都算出来再用各项指标参数加以比较,计算量就已非常大了. 在实际问题的研究过程当中,D 的维数往往远远高于 100. 例如,在利用生物芯片来进行药物设计和癌症诊断时,其产生的有效特征维数往往在 10 000 左右. 而实际需要选取的优化特征组的特征数量 d 是未知的[10-11]. 因而,寻找可行的特征选择算法已逐渐成为国际上研究的热点,这类算法也是数据挖掘的主要理论课题之一.

目前国际上进行该项研究的着眼点主要放在选择优化特征集合所需要的两个主要步骤上. 要确定关键的变量组,即优化的特征子集,首先必须确定进行搜索所需要的策略,其次,需要确定特征的评价准则来评价所选择的特征子集的性能[12]. 因而,可以把特征选择方法从这两个方面进行分类.

2.1 按搜索策略划分特征选择算法

根据算法进行特征选择所用的搜索策略,可以把特征选择算法分为采用全局最优搜索策略、随机搜索策略和启发式搜索策略 3 类[12-14].

1) 采用全局最优搜索策略的特征选择算法. 到目前为止,唯一应用全局最优搜索策略得到最优结果的特征方法是"分支定界"(Branch and Bound) 算

法[15-17]. 该算法的主要思路是:定义一个评价准则函数,该评价准则函数必须满足单调性条件,也就是对两个特征子集 S^1 和 S^2 而言,如果 S^1 是 S^2 的子集,那么 S^1 所对应的评价函数值必须要小于 S^2 所对应的评价函数值. 在定义了该评价函数的前提下,该算法对最终特征子集的选择过程可以用一棵树来描述. 树根是所有特征的集合,从树根往下,在树的每级每支都舍弃一个特征,而后根据可分性判据值和事先定义的最佳特征子集的特征数目,搜索满足要求的特征子集. 这个过程中可使用各种树搜索算法,例如"自上而下"搜索方法. 使用不同的搜索算法可以提高该方法的搜索效率,加快运行速度,相比于耗尽搜索方法大大节约计算时间,但是从统计的角度而言,它的运算时间数量级与耗尽搜索仍然相差不远.

采用这种搜索策略的特征选择算法,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特征子集. 但在这类算法当中至少存在 3 个问题:1) 由于该算法无法对所有的特征依据其重要性进行排序,如何事先确定优化特征子集当中特征的数目就成了一个很大的问题,该问题影响到算法的求解规模,且与实际应用紧密联系. 2) 合乎问题要求的满足单调性的可分性判据难以设计. 这里的合乎问题要求指能够让最后选择出的特征子集符合决策的要求,取得较高的决策正确率. 3) 当处理高维度多类问题时,算法要运行多次,算法运算效率低下的问题将非常明显.

2) 采用随机搜索策略的特征选择算法. 特征选择本质上是一个组合优化问题,求解这类问题可采用非全局最优目标搜索方法,其实现靠带有一定智能的随机搜索策略. 它在计算过程中把特征选择问题和模拟退火算法、禁忌搜索算法、遗传算法[18]、或随机重采样(Bootstrap)过程[19-20]结合,以概率推理和采样过程作为算法基础. 基于对分类估计的有效性,在算法运行中对每个特征赋以一定的权重,而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上述阈值,它就被选中作为重要特征来训练分类器或作为结果输出. 这里对特征重要性的评价既可以采取特征的统计得分,也可以采用特征对分类器的贡献率[20].

遗传算法在这一领域的应用最为广泛[12,21-24]. W. Siedlechi 等学者提出早期的基于遗传算法和 k 近邻分类器的特征选择方法. 然后是 J. Yang 等学者又提出了使用遗传算法结合人工神经网络分类器进行特征选择的方法. 在后者的方法中使用基于等级排序的遗传算法选择策略. 在他们的等级选择方法中预置一个参数 $p \in (0.5,1)$,它被设置为最高效的特征被选择的概率,被排在第 k 位的特征的概率为 $p(1-p)^{k-1}$. 他们使用几个标准的实际模式识别问题来测试算法,并得到较好结果. 但是他们在适应度函数当中用到测试集精度,这给最终的分类器构造带来不可避免的偏置,也就是所选择出的特征只能有效表征出现过的数据,在新出现的数据上这些特征的性能无法得到保证. L. H. Chiang 等学者引入用特征得分比较它们之间相对重要性的方法,使用二进编码和标准的遗传操作数来表述特征选择问题. 采用根据基于训练集上分类性能制定的特征评价函数,结合几个化工过程关键变量辨识问题,他们认为遗传算法在大量计算的基础上得到较鲁棒的结果.

从解决问题的效果来看,该方法把分类性能引入作为特征的评价准则,在此基础上对所有的特征进行排序,可以得到一个较好的应用结果. 但是这类方法同样也存在着一些问题:1) 大量的时间消耗. 它的运行时间随着数据集变量的增加呈指数增长,对于特征很多的数据集(如生物芯片数据等)时间消耗特别大. 2) 如果在该类算法中采用的是统计得分的方式,那么仅能对所有特征的重要性进行排序,很难确定一个优化的特征子集. 而采用类似于 L. H. Chiang 等人所用的方法,那么在算法运行的过程中就必须指定优化特征子集的特征数目的上限,算法的复杂度与该上限呈指数关系,而在实际应用中该上限事先难以确定. 上述两个问题在特征选择算法结合 one vs. one 的多类分类算法中体现尤为突出.

3) 采用启发式搜索策略的特征选择算法主要有以下 8 种.

(1) 单独最优特征组合. 该方法依靠计算各特征单独使用时的判据值对特征加以排队. 取前 d 个特征作为满足条件的特征组. 这种方法仅当单个特征的判据值满足加和性或乘性条件的时候才能选择出一组最优的特征. 例如在两类问题中,当两类都是正态分布情况,且各个特征间统计独立的时候,用 Mahalanobis 距离作为可分性判据,则可以达到这样的效果. 但是特征间具有这种关系仅仅是极少数情况,大多数状况下,该算法甚至可能取到最差的特征组合[3]. 在很多情况下,该方法可以用来去掉一些不重要的变量,例如对所有变量排序,而后去掉排在后面的一定数目的变量. 由于特征排序采用的判据

计算较为简单，因此在很大程度上可以很快地缩减特征选择的范围. 是一种较好的特征预选方法.

(2) 序列前向选择方法(Sequential Forward Selection，SFS)，也称为集合增加法. 它是一种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，每次向特征集合中增加一个特征，当所需要的特征集合达到要求时所得到的特征集合作为算法运行的结果. 该过程可以描述为：设所有的特征集合为 Q，假设有一个已有 d_1 个特征的特征集 X_{d_1}，对每一个未入选特征 ξ_j(即 $Q-X_{d_1}$ 中的特征)计算其准则函数 $J_j=J(X_{d_1}+\xi_j)$. 选择使 J_j 最大的那个特征，并把它加入到集合 X_{d_1} 中. 实际上，在算法的每一步，都选择一个特征加入到当前集合，使得特征选择准则最大. 当最佳改进使特征集性能变坏或达到最大允许的特征个数的时候，该算法认为已经选择出最佳特征子集. 该算法的运算量相对较小，但是特征之间的统计相关性没有得到充分考虑. 从这个角度出发的搜索方式仅能适合一小部分满足特殊条件的特征集合. 例如算法第一步选出的必然是使准则函数最大的一个特征，而后来每步选出的都是对前一个特征集合作为最佳补充的一个特征[25]. 在实际过程中，最佳特征集合极有可能并不包括单独贡献率(准则函数值)最大的那个特征，仅仅只是一些单独贡献率极为普通的特征组合. 在该算法中每步都可能出现这样的现象.

(3) 广义序列前向选择方法(Generalized Sequential Forward Selection，GSBS). 该方法是SFS算法的加速方法，它可以根据准则函数一次性向特征集合中增加 r 个特征. 也就是在没有入选优化特征子集的剩余特征集中，寻找一个规模为 r 的小特征集 Y_r，使得 $J(X_{d_1}+Y_r)$ 最大. 该方法相对于SFS在特征统计相关性上要稍好些，但是计算量相对SFS增大许多，且在SFS中出现的问题依旧难以避免.

(4) 序列后向选择方法(Sequential Backward Selection，SBS). 该方法是一种自上而下的方法. 该方法在运行之初假定整个特征集合就是所需要的优化特征集. 而后在算法的每步运行过程中删除一个对准则函数无贡献的特征，直到剩余特征个数符合集合基数要求. 该方法在一个较大的变量集上计算准则函数 J，所以该算法相对于SFS计算量要大. 该方法的优势在于充分考虑特征之间的统计相关特性，因而在采用同样合理的准则函数的时候，它的实际计算性能和算法的鲁棒性要大大优于SFS算法.

(5) 广义序列后向选择方法(Generalized Sequential Backward Selection，GSBS). 该方法是SBS算法的加速算法，它根据准则函数在算法的每个循环当中，一次性删除一定个数的无用特征. 它是一种可应用于实际过程的快速特征选择方法. 它的优点在于速度较快，性能相对较好. 不足之处在于有的时候，特征消除操作进行太快，容易丢失重要的变量，导致找不到最优的特征组.

(6) 增 l 去 r 选择方法. 这种方法允许在特征选择过程中进行回溯，如果 $l>r$，则该算法是自下而上的方法. 用SFS方法将 l 个特征加入到当前特征集中，然后再用SBS方法删除 r 个最差的特征. 这种方法消除嵌套问题，因为某一步获得的特征集不一定是下一步特征集的子集. 如果 $l<r$，则算法为自上而下的方法. 从一个完全特征集开始，依次删除 r 个特征，再增加 l 个特征直到获得满足要求个数的特征. 该方法实际上是SBS方法和SFS方法的一种折衷，它的运算速度要比SBS快，运算效果要比SFS好.

(7) 广义增 l 去 r 选择方法. 该方法是在增 l 去 r 选择方法的基础上，用GSFS和GSBS分别代替SFS和SBS. 前面所有讨论过的算法甚至可以看作是它的特例算法. 因而它包含极其广泛的理论意义. 但操作较为复杂，难以制定实际规则加以利用.

(8) 浮动搜索方法. 该方法改变上述一系列算法固定 l、r 的基本做法. 采用浮动的步长，也就是在选择算法的不同步骤，可以采用不同的 l、r. 实际的每轮的 l、r 可以根据特征的统计特点来制定. 这是一种非常实用的改良机制.

上述方法中，一般认为采用浮动广义后向选择方法(Floating Generalized Sequential Backward Selection，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特点，又用浮动方法保证算法运行的快速稳定性[26].

综上所述，根据合理的启发式规则可以设计出非常实用的次优搜索方法应用于特征选择算法. 该类算法并不检查每个特征组合，但是它可以估计一组潜在、有用的特征组合，甚至可以根据所制定的启发式规则对所有特征进行排序. 在合理设计规则的作用下，实际应用中这类算法甚至能够达到和前两种搜索策略类似的效果，且具有运算速度快等特点. 文献[26]～[28]也从运算量和运行性能等几个角度出发，分析几个实际问题，展示这类算法在很多状况下可以作为前两种搜索算法的最佳替代方法.

2.2 按特征集合评价策略划分特征选择算法

从特征集合的评价策略上来分,特征选择方法大致可以分成两类:滤波器方法(Filter)及嵌入式方法(Wrapper).这两者的区别在于优化特征子集的评价是否用到在决策机器构造过程中所使用的学习算法.如果用到,那么就是 Wrapper 方法,否则就属于 Filter 方法[12-13].

1)基于滤波(Filter)评价策略的特征选择算法.滤波特征选择方法是一种计算效率较高的方法.这类方法使用合适的准则来快速评价特征的好坏.这类特征选择方法可以参照文献[29]~[31].在这些文献中,设计这些准则既可以用来削减特征之间的相关性[32],也可以用来削减特征之间的互信息[33].除了这些准则之外,还可以使用其它一些较为简单的信息统计准则[10,34-35].目前用的最多的是概率距离和相关测量法[30]、类间和类内距离测量法[31,36]、信息熵法[37]、决策树滤波方法[38]等.这些方法存在的一个主要问题是它并不能保证选择出一个规模较小的优化特征子集,尤其是当特征和分类器关联较大时.因而,即使该类方法可以找到一个满足条件的优化子集,该子集的规模也会较大,在其中会包含一些明显的噪声特征,给寻找关键性的特征和产生类区别的源头带来较大阻碍.该方法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范围,用来作为特征的预选器非常好.

2)基于嵌入式(Wrapper)评价策略的特征选择算法.嵌入式方法和所使用的分类器有很大关系.它在筛选特征的过程当中直接用所选特征来训练分类器,根据这个分类器在验证集上的表现来评价所选择的特征.该方法在速度上比滤波方法慢,但其所选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和精简诊断决策机器的结构.目前此类方法是特征选择研究领域的热点,关于这类方法的描述有许多相关文献.在文献[39]中,作者设计用决策树来进行特征选择的 Wrapper 方法,文中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.文献[22]中,Fisher 判别分析结合遗传算法被用来在化工故障过程中用来辨识关键变量,取得不错效果.文献[34]结合正态极大似然模型来进行特征选择和分类,并给出满意的实验结果.在文献[40]中,作者用遗传算法结合人工神经网络进行同样尝试.类似的文献还有文献[41]~[43]等.在以上这些 Wrapper 方法中,作者都是直接通过分类器的分类性能来评价特征的可用性.这类方法实际上是 Wrapper 方法和随机搜索策略的结合,其结果较好,但是运算的时间较长.在文献[29]、[32]、[44]中用启发式搜索策略(SBS,SFS, FSFS)和分类器性能评价准则相结合来评价所选的特征,也取得不错的效果,相对于使用随机搜索策略的方法,节约不少时间.但是该类算法的实用性依然是一个值得关注的问题.这个过程的时间主要消耗在成千上万次分类器的训练和这些分类器在测试集上的性能验证上.它的主要问题存在于评价策略的直接性上.间接的基于 Wrapper 的评价准则可以极大的改善算法的运行效率,根据这种从分类器本身所派生出来的启发式的评价准则,可以极大地改善算法效率.文献[45]中提出一种非常经典的 Wrapper 方法和启发式搜索准则(SBS)相结合的特征选择方法.作者从支持向量机本身的理论出发,根据分类器的决策系数来决定特征的贡献(评价准则),最终在一个 7 129 维的特征集中获取一个 64 维的优化特征子集,且运行时间也是可以接受的.类似的采用神经网络实现的特征选择算法在文献[43]中也已提出,并取得非常不错的应用效果.还有一些较特殊的应用启发式搜索策略的特征选择算法,例如在文献[46]当中,作者根据支持向量机本身的界理论,提出将特征选择问题转化为一个用整数规划求解使目标函数全局最大化的参数选择问题,作者认为这种方法甚至可以用来解决非线性特征选择问题.在这种方法当中的启发式规则实际上是各特征对目标函数的梯度值.但是这样又引起其它问题,例如梯度相关参数的设置和全局优化函数的合理性等,该方法始终未得到推广应用.

综合以上所提到的搜索策略和特征集的评价准则,可以看到采用基于启发式搜索策略的 Wrapper 方法是较实用的进行特征选择操作的热点研究方向.

可以用来评价所选择出的一组特征的方法主要包括:1)该特征组是否具有可解释的实际意义;2)使用该特征组训练得到的分类器在训练集、验证集或者测试集上是否具有较好的统计判别性能.这些统计性能测试方法主要包括交叉验证,自助法,两类情况下的 ROC 曲线[9,47],更加严格的性能测试,还包括扰动测试实验等[48].

3 基于支持向量机的应用启发式搜索策略的特征选择方法

采用基于启发式搜索策略的 Wrapper 方法是

目前的研究热点所在.采用这类方法所面临的一个很大问题是必须要选取一类非常可靠的分类算法来做为整个特征选择方法的基础.支持向量机(support vector machine)是近年发展起来的新型的通用知识发现方法,对于一般规模数据量的、含有噪声模式且缺乏统一理论的领域非常适用[49-50].它本身是一种有监督的学习算法.文献[22]、[31]、[51]等多篇文献表明该方法具有相当的普适性.

虽然基于支持向量机的特征选择策略有许多[41,52],但从本文观点来看,基于启发式搜索策略的支持向量机的 Wrapper 方法是其中最值得研究的部分.从搜索策略看,它主要包括 RFR(Recursive Feature Replacement)和 RFE(Recursive Feature Elimination)2 种方法.

基于支持向量机的 RFR 方法发表于文献[53]中,该方法是基于 SFS 的自底向上的方法.它以特征集的交叉验证错误率为评价指标,首先初始化所需要的特征子集为空集,而后将使得特征集交叉验证率为最小的新的特征加入到这一集合当中,直到将所有特征逐一排序,在实验中取得不错结果.但从理论角度上看,该方法采用 SFS 搜索方法为主线,不能对特征间的统计关联加以充分考虑.以交叉验证率为指标来搜索大量的特征组合,计算量也较大.在该方法中仅仅考虑特征之间最多仅有线性关联的情况,因而该方法还有待改进.

SVM-RFE(Recursive feature selection based on support vector machine)方法,也称为基于支持向量机的回归特征消去方法,是本文主要讨论的算法对象之一.该算法由法国学者 I. Guyon 在文献[45]当中提出.该算法认为使用 RFE 算法可以保证在特征排序的过程中保留优化特征子集.在对特征进行排序时,该方法使用支持向量机的判别函数中的信息来实现.

RFE-回归特征消去方法,它是一个循环的过程,在这个过程的每步都包含以下 3 个步骤:1)用当前数据集训练分类器,根据所得分类器获得所使用特征的相关信息,例如,在线性核支持向量机当中,这些相关信息为每个特征的权重.2)根据事先制定的规则,计算所有特征的排序准则分数.3)在当前数据集中移除对应于最小排序准则分数的特征.该循环过程执行到特征集中剩余最后一个变量时结束.算法执行结果为一列按照特征重要性排序的特征序号列表.可以看出,这个迭代的过程实际上是一个序列后向选择(SBS)的过程,它在整个循环中先是去除与判别不相关的特征,保留对判别相对重要的优化特征子集,因而可以达到优化特征子集选择,提高判别精度的目的.与此同时,该方法使用的是一个间接的启发性排序准则,在算法速率上得到一定保障.在文献[45]当中,原始的 SVM-RFE 算法用的是 SBS 和 GSBS 方法,因而从运算速率和保证性能两方面来说,还存在着明显缺憾.文献[26]、[54]提出该算法的基于浮动搜索策略的加速算法,对这方面进行一些改进.当特征间存在的为非线性关联的情况下,可以将 SVM-RFE 用非线性核支持向量机加以实现,可以得到较为满意的结果[55,57].该方法当中存在的另一个问题是如何处理冗余特征,过多的冗余特征将使所得到的优化特征子集的规模更大,而判别能力下降.因而在该算法当中嵌入惩罚冗余特征的指标非常重要[56].

4 结束语

本文从不同角度对特征选择算法进行分类,概述特征选择发展的各个分支方向及发展态势,指出理论研究和实际应用中所存在的困难和一些亟待解决的问题.然后从实用的角度出发,结合机器学习的观点,探讨应用支持向量机技术来进行特征选择的研究思路,指出该方法未来的一些发展方向.

参考文献

[1] de Sa Marques J P. Pattern Recognition Concepts, Methods and Applications. Berlin, Germany: Springer-Verlag, 2002

[2] Ganeshanandam S, Krzanowski W J. On Selecting Variables and Assessing Their Performance in Linear Discriminant Analysis. Australian Journal of Statistics, 1989, 31(3): 433-447

[3] Bian Zhaoqi, Zhang Xuegong. Pattern Recognition. 2nd Edition. Beijing, China: Tsinghua University Press, 2000 (in Chinese)
(边肇祺,张学工. 模式识别. 第 2 版. 北京:清华大学出版社, 2000)

[4] Theodoridis S, Koutroumbas K. Pattern Recognition. 2nd Edition. New York, USA: Elsevier, 2003

[5] Dougherty E R. Small Sample Issues for Microarray-Based Classification. Comparative and Functional Genomics, 2001, 2(1): 28-34

[6] Dougherty E R, Shmulevich I, Bittner M L. Genomic Signal Processing: The Salient Issues. EURASIP Journal on Applied Signal Processing, 2004, 4(1): 146-153

[7] Kim S, Dougherty E R, Barrera J, *et al*. Strong Feature Sets from Small Samples. Journal of Computational Biology, 2002, 9(1): 127-146

[8] Hastie T, Tibshirani R, Friedman J. The Elements of Statisti-

cal Learning: Data Mining, Inference, and Prediction. New York, USA: Springer-Verlag, 2001

[9] Webb R A. Statistical Pattern Recognition. New York, USA: John Wiley & Son, 2002

[10] Dudoit S, Fridlyand J, Speed T P. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association, 2002, 97(457): 77-87

[11] Adam B L, Vlahou A, Semmes O J, *et al*. Proteomic Approaches to Biomarker Discovery in Prostate and Bladder Cancers. Proteomics, 2001, 1(10): 1264-1270

[12] Sun Z H, Bebis G, Miller R. Object Detection Using Feature Subset Selection. Pattern Recognition, 2004, 37(11): 2165-2176

[13] Jain A K, Duin R D W, Mao J C. Statistical Pattern Recognition: A Review. IEEE Trans on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37

[14] Kudo M, Sklansky J. Comparison of Algorithms That Select Features for Pattern Classifiers. Pattern Recognition, 2000, 33(1): 25-41

[15] Chen Xuewen. An Improved Branch and Bound Algorithm for Feature Selection. Pattern Recognition Letters, 2003, 24(12): 1925-1933

[16] Fukunaga K, Narendra P M. A Branch and Bound Algorithm for Computing k-Nearest Neighbors. IEEE Trans on Computers, 1975, 24(7): 750-753

[17] Hamamoto Y, Uchimura S, Matsuura Y, *et al*. Evaluation of the Branch and Bound Algorithm for Feature Selection. Pattern Recognition Letters, 1990, 11(7): 453-456

[18] Wang Ling. Intelligent Optimization Algorithms with Applications. Beijing, China: Tsinghua University Press, 2004 (in Chinese)
(王 凌.智能优化算法及其应用.北京:清华大学出版社, 2004)

[19] Tsymbal A, Puuronen S. Ensemble Feature Selection with the Simple Bayesian Classification. Information Fusion, 2003, 4(2): 87-100

[20] Wu B L, Abbott T, Fishman D, *et al*. Comparison of Statistical Methods for Classification of Ovarian Cancer Using Mass Spectrometry Data. Bioinformatics, 2003, 19(13): 1636-1643

[21] Yang J, Honavar V. Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems, 1998, 13(2): 44-49

[22] Chiang L H, Pell R J. Genetic Algorithms Combined with Discriminant Analysis for Key Variable Identification. Journal of Process Control, 2004, 14(2): 143-155

[23] Siedlecki W, Sklansky J. A Note on Genetic Algorithms for Large Scale Feature Selection. Pattern Recognition Letters, 1989, 10(11): 335-347

[24] Peng Sihua, Xu Qianghua, Ling Xuefeng. Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithms and Support Vector Machines. FEBS Letters, 2003, 555(2): 358-362

[25] Mao K Z. Fast Orthogonal Forward Selection Algorithm for Feature Subset Selection. IEEE Trans on Neural Networks, 2002, 13(5): 1218-1224

[26] Furlanello C, Serafini M, Merler S, *et al*. An Accelerated Procedure for Recursive Feature Ranking on Microarray Data. Neural Networks, 2003, 16(5/6): 641-648

[27] Somol P, Pudil P, Novovičová J, *et al*. Adaptive Floating Search Methods in Feature Selection. Pattern Recognition Letters, 1999, 20(11/12/13): 1157-1163

[28] Pudil P, Novovicova J, Kittler J. Floating Search Methods in Feature Selection. Pattern Recognition Letters, 1994, 15(11): 1119-1125

[29] Inza I, Larranaga P, Blanco R, *et al*. Filter Versus Wrapper Gene Selection Approaches in DNA Microarray Domains. Artificial Intelligence in Medicine, 2004, 31(2): 91-103

[30] Zhou Xiaobo, Wang Xiaodong, Dougherty E R. Nonlinear-Probit Gene Classification Using Mutual-Information and Wavelet-Based Feature Selection. Biological Systems, 2004, 12(3): 371-386

[31] Furey T S, Cristianini N, Duffy N, *et al*. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics, 2000, 16(10): 906-914

[32] Zhou Xiaobo, Wang Xiaodong, Dougherty E R. Gene Selection Using Logistic Regressions Based on AIC, BIC and MDL Criteria. Journal of New Mathematics and Natural Computation, 2005, 1(1): 129-145

[33] Zhou Xiaobo, Wang Xiaodong, Dougherty E R. Construction of Genomic Networks Using Mutual Information Clustering and Reversible-Jump Markov Chain Monte Carlo Predictor Design. Signal Processing, 2003, 83(4): 745-761

[34] Tabus I, Astola J. On the Use of MDL Principle in Gene Expression Prediction. EURASIP Journal of Applied Signal Processing, 2001, 4: 297-303

[35] Liu Huiqing, Li Jinyan, Wong L. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics, 2002, 13: 51-60

[36] Michael M, Lin W C. Experimental Study of Information Measure and Inter-Intra Class Distance Ratios on Feature Selection and Orderings. IEEE Trans on System, Man, and Cybernetics, 1973, 3(2): 172-181

[37] Sindhwani V, Rakshit S, Deodhare D, *et al*. Feature Selection in MLPs and SVMs Based on Maximum Output Information. IEEE Trans on Neural Networks, 2004, 15(4): 937-948

[38] Haering N, Lobo N D V. Feature and Classification Methods to Locate Deciduous Trees in Images. Computer Vision and Image Understanding, 1999, 75(1/2): 133-149

[39] Hsu W H. Genetic Wrappers for Feature Selection in Decision Tree Induction and Variable Ordering in Bayesian Network Structure Learning. Information Sciences, 2004, 163(1/2/3): 103-122

[40] Li L, Weinberg C R, Darden T A, *et al*. Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method.

Bioinformatics, 2001, 17(12): 1131-1142

[41] Shima K, Todoriki M, Suzuki A. SVM-Based Feature Selection of Latent Semantic Features. Pattern Recognition Letters, 2004, 25(9): 1051-1057

[42] Jack L B, Nandi A K. Fault Detection Using Support Vector Machines and Artificial Neural Networks, Augmented by Genetic Algorithms. Mechanical Systems and Signal Processing, 2002, 16(2/3): 373-390

[43] Verikas A, Bacauskiene M. Feature Selection with Neural Networks. Pattern Recognition Letters, 2002, 23(11): 1323-1335

[44] Xiong Momiao, Fang Xiangzhong, Zhao Jinying. Biomarker Identification by Feature Wrappers. Genome Research, 2001, 11(11): 1878-1887

[45] Guyon I, Weston J, Barnhill S, *et al*. Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, 2002, 46(1/2/3): 389-422

[46] Weston J, Mukherjee S, Chapelle O, *et al*. Feature Selection for SVMs // Solla S A, Leen T K, Muller K R, eds. Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2001, 13: 668-674

[47] Perner P, Apte C. Empirical Evaluation of Feature Subset Selection Based on a Real-World Data Set // Proc of the 4th European Conference on Principles of Data Mining and Knowledge Discovery. London, UK: Springer-Verlag, 2000: 575-580

[48] Zhang Xuegong, Wong W H. Recursive Sample Classification and Gene Selection Based on SVM: Method and Software Description. Technical Report, Boston, USA: Harvard School of Public Health. Department of Biostatistics, 2001

[49] Brown M P S, Grundy W N, Lin D, *et al*. Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. Proc of the National Academy of Science, 2000, 97(1): 262-267

[50] Barzilay O, Brailovsky V L. On Domain Knowledge and Feature Selection Using a Support Vector Machine. Pattern Recognition Letters, 1999, 20(5): 475-484

[51] Fortuna J, Capson D. Improved Support Vector Classification Using PCA and ICA Feature Space Modification. Pattern Recognition, 2004, 37(6): 1117-1129

[52] Simek K, Fujarewicz K, Swierniak A, *et al*. Using SVD and SVM Methods for Selection, Classification, Clustering and Modeling of DNA Microarray Data. Engineering Applications of Artificial Intelligence, 2004, 17(4): 417-427

[53] Fujarewicz K, Wiench M. Selecting Differentially Expressed Genes for Colon Tumor Classification. International Journal of Applied Mathematics and Computer Science, 2003, 13(3): 327-335

[54] Mao Yong, Pi Daoying, Yu Ming, *et al*. Accelerated Recursive Feature Elimination by Support Vector Machine for Key Variable Identification. Chinese Journal of Chemical Engineering, 2006, 14(1): 65-72

[55] Mao Yong, Zhou Xiaobo, Pi Daoying, *et al*. Parameters Selection in Gene Selection Using Gaussian Kernel Support Vector Machines by Genetic Algorithm. Journal of Zhejiang University: Science B, 2005, 6(10): 961-973

[56] Li Fan, Yang Yiming. Using Recursive Classification to Discover Predictive Features // Proc of the ACM Symposium on Applied Computing. Santa Fe, New Mexico, 2005: 1054-1058

[57] Mao Yong, Zhou Xiaobo, Yin Zheng, *et al*. Gene Selection Using Recursive Feature Elimination Based on Gaussian Kernel Support Vector Machine with Adaptive Kernel Width Strategy // Proc of the 1st International Conference on Rough Sets and Knowledge Technology. Chongqing, China, 2006: 799-806