计算机科学2004Vol.31№.11

机器学习中的特征选择

张丽新　王家钦　赵雁南　杨泽红

（清华大学计算机科学与技术系　智能技术与系统国家重点实验室　北京100084）

摘　要　20世纪90年代以来，特征选择成为机器学习领域的重要研究方向，研究成果十分显著，但是也存在许多问题需要进一步研究。本文首先对特征选择和学习算法结合的三种方式进行了系统的总结；然后将一般特征选择定位为特征集合空间中的启发式搜索问题，对特征选择算法中的四个要素进行了阐述，其中重点总结了特征评估的方法；最后对特征选择的研究现状进行了回顾，分析了目前特征选择研究的不足和未来发展的方向。

关键词　特征选择，机器学习，搜索算法，评估函数

Feature Selection in Machine Learning

ZHANG Li-Xin　WANG Jia-Qin　ZHAO Yan-Nan　YANG Ze-Hong

(The State Key Laboratory of Intelligent Technology and Systems, Computer Science & Technology Department, Tsinghua University, Beijing 100084)

Abstract　Feature selection has been an important research area in machine learning since 90's of the 20th century. Great achievements have been achieved, however many problems remain to be unsolved and need further investigation. In this paper, we make systematic survey on the three combination modes of featuire selection with induction algorithm. We describe feature selection in terms of heuristic search through the space of feature sets, and discuss the foru factors in feature selection algorithms, in which the evaluation function is detailedly analyzed and discussed. Last we overview the investigation status of the feature selection, and point out the limitations of current research and challenges in future work.

Keywords　Feature selection, Machine learning, Search algorithm, Evaluation function

1　引言

所谓特征就是描述模式的属性，机器学习中的特征选择可定义为：已知一特征集，从中选择一个子集使评价标准最优[1]。以上定义可表述为：

给定一个学习算法 L，一个数据集 S，数据集 S 来自一个具有 n 个特征 $X_1, X_2, X_3, \cdots, X_n$ 的具有类别标记 Y 的符合分布的例子空间，则一个最优特征子集 X_{opt} 是使得某个评价准则 $J=J(L,S)$ 最优的特征子集。

特征选择是机器学习领域的重要问题。在一个学习算法通过训练样本对未知样本进行预测之前，必须决定哪些特征应该采用，哪些特征应该忽略。虽然在学习算法方面已经开展了大量的研究，但特征选择方面的研究则相对较少。自20世纪90年代以来，特征选择方面的研究引起机器学习领域学者前所未有的重视，主要原因有以下两个方面：1)许多学习算法的性能受到不相关或冗余特征的负面影响。已有的研究结果表明，大多数学习算法所需训练样本的数目随不相关特征的增多而急剧增加[1~4]。Langley 等的研究表明最近邻法的样本复杂度随不相关特征成指数增长，其他归纳算法也基本具有这一属性[1,2]。例如，决策树对于逻辑与概念的样本复杂度随不相关特征线性增加，但对于异或概念的样本却是呈现指数增长[1]；贝叶斯分类器虽然对不相关特征的存在不敏感，但其性却能对冗余特征的存在很敏感[2]。因此，特征选择对不同情况下的学习算法都有不可忽视的作用。选择好的特征不仅可以减小计算复杂度，提高分类准确度，而且有助于寻找更精简更易理解的算法模型。2)大规模数据处理问题的不断出现。所谓大规模，一方面指样本数目的庞大，另一方面指描述样本的特征维数高。数据挖掘的发展对大规模数据处理的研究提出了迫切的要求，如信息检索，遗传基因分析等[3,4]。正是由于上述原因，特征选择成为机器学习领域重要的研究方向，引起越来越多的机器学习领域学者的兴趣。国内外的各大研究机构如CMU，Stanford，Washington，南京大学，哈尔滨工业大学，北京工业大学等都开展相关研究[5~7]。

特征选择和学习算法是密切相关的，特征选择的结果最终由学习算法的性能来评估，因此特征选择和学习算法的结合方式也就非常重要[8]。本文系统介绍了特征选择和学习算法结合的三种结构；并从搜索寻优的角度介绍了特征选择算法的四个组成要素，重点分析了其中的特征评估；最后概述了特征选择的最新研究现状，分析了未来的研究方向。

2　特征选择和学习算法结合的三种结构

特征选择和后续学习算法的结合方式可分为嵌入式、过滤式和 Wrapper 三种。

2.1　嵌入式特征选择

在嵌入式结构中，特征选择算法本身作为组成部分嵌入到学习算法里。如某些逻辑公式学习算法是通过向公式表达式中加减特征实现的[9]。类似的加减特征操作也构成一些更复杂的逻辑概念推导的核心，只是通过不同特征组合形成更复杂的规则描述。最典型的即决策树算法，如 Quinlan 的 ID3 和 C4.5[10,11]以及 Breiman 的 CART 算法[12]等，算法在每一结点选择分类能力最强的特征，然后基于选中的特征进行子空间分割，继续此过程，直到满足终止条件，可见决策树生成的过程也就是特征选择的过程。

2.2　过滤式特征选择

过滤式特征选择的评估标准直接由数据集求得，独立于学习算法，如图1所示。最简单的过滤式特征选择于20世纪60年代早期提出[13]，该算法在特征间相互独立的假设下，研究每一特征对于分类的可分性或熵，然后选择其中评价最好的

k 个特征组合在一起。这种方法通常用在文本分类上，常和贝叶斯分类器或者最近邻分类器结合在一起。该方法虽然简单，但由于没有考虑特征间的相互作用，性能并不理想，即使在满足特征间独立的条件下，两个单独使用最好的特征组合起来也不能保证是最好的组合。

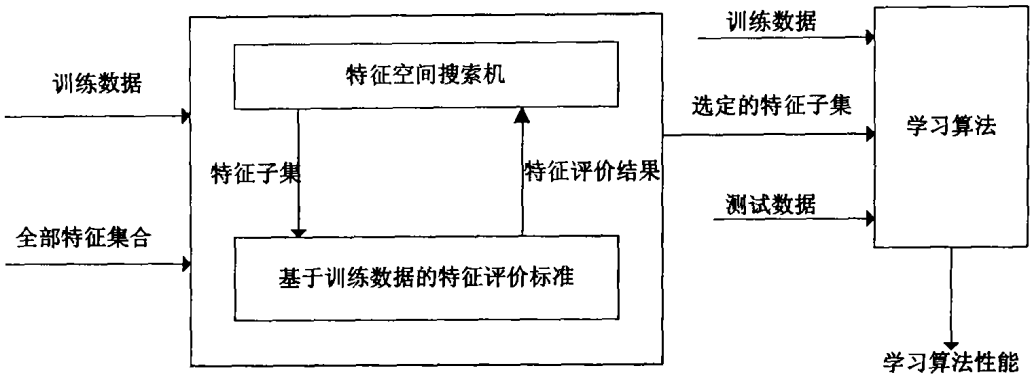

图1　过滤式(Filter)特征选择一般流程

另一种过滤式特征处理方法是从原始特征集通过映射或变换构造出新特征，通常被称为特征提取。主成分分析是这类方法中最著名的算法，该算法对许多学习任务都可以较好降维，但是特征的理解性很差，因为即使简单的线性组合也会使构造出的特征难以理解，而在很多情况下，特征的可理解性是很重要的。

2.3　Wrapper 特征选择

Wrapper 特征选择算法最早由 John 等在1994年提出[14]，如图2所示。该算法的主要思想是：和学习算法无关的过滤式特征评价会和后续的分类算法产生较大的偏差，而学习算法基于所选特征子集的性能可以作为更好的特征评价标准。因此在 Wrapper 特征选择中将学习算法的性能作为特征选择的评估标准。

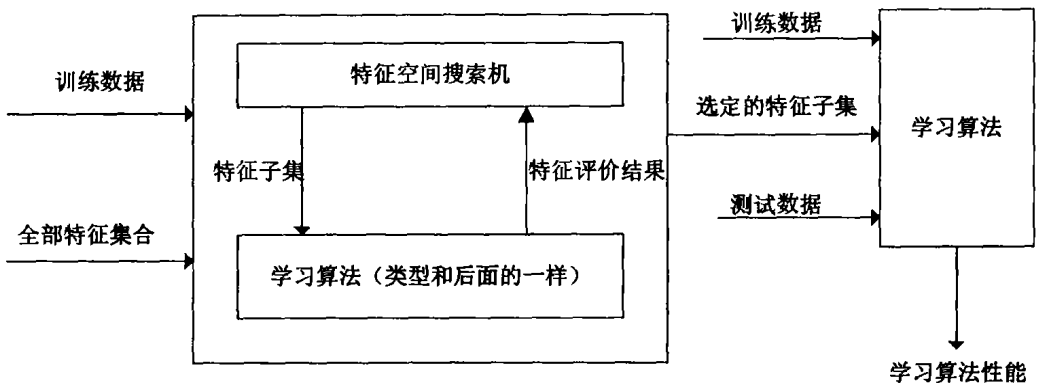

图2　Wrapper 特征选择一般流程

Wrapper 特征选择算法中用以评估特征的学习算法是没有限制的。John 等选用决策树[14]，Aha 等将最近邻法 IB1和特征选择算法相结合对云图进行分类研究[15]，Provan，Inza 等则利用贝叶斯网络性能指导贪心的前向搜索算法[16,17]。

由于采用学习算法的性能作为特征评估标准，Wrapper 特征选择算法比过滤式特征选择算法准确率高，但算法效率较低。因此一些研究者努力寻找使评价过程加速的方法。Caruana 等提出一种加速决策树的方法[18]，即在特征选择过程中大量减少决策树分支的数目。Moore 等通过减少评估特征阶段的分类器的训练样本来提高特征选择的速度[19]。Wrapper 方法的另一个缺点是过适应问题，但该问题主要发生在训练数据规模较小的情况[8]。

3　特征选择作为搜索问题的四要素

一般而言，特征选择可以看作一个搜索寻优问题。对大小为 n 的特征集合，搜索空间由 2^n 种特征选择的可能状态构成。Davies 等证明最小特征子集的搜索是一个 NP 问题[20]，即除了穷尽式搜索，不能保证找到最优解。但实际应用中，当特征数目较多的时候，穷尽式特征选择因计算量太大而无法应用，因此人们致力于用启发式搜索算法寻找次优解。一般特征选择算法必须确定以下四个方面：1）搜索起点；2）搜索策略；3）特征评估函数；4）终止条件。本文重点介绍特征评估函数，其余三方面则概要介绍，感兴趣的读者可参阅文[21]。有些特征选择算法只包含以上四个要素的部分内容，比如特征加权后选择前 m 个特征的特征选择算法，就只涉及特征评估和终止条件两方面。

3.1　搜索起点、搜索策略和终止条件

搜索起点是算法开始搜索的状态点，搜索起点的选择对搜索策略有重要影响。如果搜索起点为空集，通常就要逐个地向集合里加入特征，即所谓的前向搜索；如果搜索起点为全集，就要不断地删减特征，即所谓的后向搜索；如果搜索从特征空间的中间结点开始，那么搜索策略通常就是随机的或者启发式的搜索。

根据搜索方向、搜索策略，可以分为前向、后向和双向三种。对于小规模的特征集合，可采用穷尽式搜索求得最优子集。对于中等规模的特征集合，当评估函数对特征维数满足单调性时，可采用 Narendra 等提出的分支界限法(BB)求解最优特征子集[22]。但实际问题中，评估函数通常不具备单调性，同时 BB 算法的算法复杂度与特征个数之间是指数关系，在 n 较大时由于计算量太大而无法应用。因此人们一直致力于寻找能得到较好次优解的搜索算法。特征子集搜索算法大致可分为顺序搜索和随机搜索两类。顺序搜索算法采用顺序地向

解集中加减特征逐步扩展搜索，如顺序前向搜索，顺序后向搜索，以及广义的顺序前向搜索和广义的顺序后向搜索等。该类算法的缺点是，特征一旦被加入或删除，以后便不会改变，因此容易陷入局部极值。为克服此缺点，出现了增 l 减 r 法，即搜索方向不再是单向加或者减，可以根据评估函数灵活的浮动，其问题在于 l 和 r 的大小难以确定[21]。Pudil 等提出了顺序浮动前向搜索和顺序浮动后向搜索算法[23]，算法变固定的增 l 减 r 法为浮动的，减少了不必要的回溯并在需要时增加回溯的深度，Somol 等进一步提出了自适应浮动搜索算法[24]，根据当前特征子集的大小和目标特征子集的大小来控制搜索空间的大小，这种方法减小了陷入局部极值的可能性。随机搜索算法包括遗传算法、模拟退火和集束式搜索(Beam Search)等[21]。遗传算法在特征选择中的应用研究很多，并且显示出良好的性能[25,26]。关于各搜索算法的优劣并没有一致的意见，但根据 Jain[3]和 Kudo[27]的实验，自适应浮动搜索算法和遗传算法是众多搜索中性能较好的算法。

特征选择的终止条件有最大运行次数，性能不改进的循环运行次数，找到满足评价函数的特征集合等[1]。

3.2 特征评估函数

特征评估函数和特征相关分析关系密切，从某种意义上说，特征的评估就是要评价特征或者特征子集和目标函数的相关性。以下首先在理论上给出了特征相关分析的概念，然后具体介绍并分析比较了几种特征评估方法。

3.2.1 *相关分析的概念* 人类在面对多种特征选择时，可以迅速定位到那些和目标函数相关的特征而忽略其他无关或不重要的特征。在机器学习中，要选择和目标函数相关的特征，首先必需对特征相关性进行定义。John 等以及 Blum 等对特征的相关性分析进行了详细阐述[28,29]。下面给出几种典型的定义并对其进行分析。在以下定义中，A 和 B 表示样本集中的样本，I 表示实例空间，S 表示样本集合，C 表示目标函数，$C(A)$表示样本 A 所属的类别(或称目标函数的值)。$X_i(A)$表示样本 A 的特征 i 的值。

定义1(特征和目标函数相关) 如果实例空间内存在样本 A 和 B，A 和 B 仅有属性 X_i 不同而 $C(A)\neq C(B)$，则称特征 X_i 和目标函数 C 相关。

这个定义的贡献在于首次从理论上给出一个严格定义，但实际应用中由于训练数据规模有限，很多情况下根本不存在只有一个属性值不同的样本对，因而无法根据这一定义确认相关特征。其次，如果存在两个特征冗余但和目标相关的情况，这个定义会将这两个特征均视为无关特征。

为了弥补上述定义的不足，John 等给出了关于样本集的强/弱相关的定义[28]，该定义针对具体的数据集而不是全体实例空间。

定义2(关于样本集强相关和弱相关) 如果在样本集合 S 中存在样本 A 和 B，仅有属性 X_i 和类别不同，则说特征 X_i 对于样本集强相关。类似地，如果在舍弃一些特征后，特征 X_i 变为强相关特征，则称特征 X_i 关于样本集弱相关。

关于样本集强相关的定义和定义1类似，只是用“样本集”代替了定义1中的“实例空间”，关于样本集弱相关定义扩充了原有的相关性定义，在实际应用中，可能某特征并不决定目标函数，但是强相关特征去掉时，该特征就起决定作用。

以上定义都是独立于学习算法且从单个特征和目标函数或者样本集的关系进行定义。但单个特征和目标函数或者样本集相关，并不能保证该特征就会对特定的分类算法有用，也不能保证多个相关的特征组成的集合就会使学习算法取得好的性能。Caruana 等考虑了特征和算法性能间的关系，给出了关于算法增益性相关的定义[30]。

定义3(关于算法增益性相关) 给定一个数据集 S，一个学习算法 L，一个特征集合 A，如果该算法 L 使用$\{X_i\}\cup A$ 所得分类的准确率比使用 A 的准确率高，则特征 X_i 对于算法 L 相对于 A 是增益性相关的。

3.2.2 *具体特征评估方法* 实际应用中很少直接应用特征相关的概念进行特征选择，但关于目标函数或样本集相关的概念是过滤式特征评估标准的理论基础，而关于算法增益性相关的概念可以看作 Wrapper 特征选择的一个理论诠释。在相关性分析的基础上，研究者提出各种更具有操作性的特征评估标准，特征的评价标准大致可分为：类内类间距离，不一致度，基于信息熵的评估标准，ReliefF 评估和学习算法性能相关的评估[12,28,29,31~33]。和学习算法性能相关的评估标准又分为：直接用学习算法性能作为特征的评价标准以及间接应用学习算法性能的评估标准[34,35]。下面将具体予以介绍。

·距离度量 是利用类内距离和类间距离的概念选择特征子集使得同类的样本分布密集而不同类的样本远离。一般需要根据样本集估计出样本均值、类间离散度矩阵和类内离散度矩阵，进而得到类内和类间距离度量函数。另外 Bhattacharyya 距离和 Chernoff 界限等也可以用来表示特征空间中两类的距离[21]。

·不一致度 是 Almuallim 等在 FOCUS 特征选择算法中提出的特征评估标准[36]。定义各特征的一种取值组合为一种模式，如果存在 l 个特征 $f_1,f_2,\cdots,f_l$，分别对应 $n_{f1},n_{f2},\cdots,n_{fl}$种取值，则最多存在的模式为 $n_{f1}*n_{f2}*\cdots*n_{fl}$种。在某种模式下如果两个样本除了类别外其余属性都相同，则说这两个样本在这种模式下是不一致的。对一个固定的样本集来说，某种模式下的不一致个数定义为在这种模式下所有样本个数和其中比例最大的类别的样本个数的差。不一致度为所有模式下不一致个数的和占全部样本个数的百分比。不一致度适用于离散特征，对连续特征，需要首先进行离散化。而且由于采用不一致度评估的特征选择算法需要考虑各种特征子集组合下，各种特征的值的组合形式，因此对于高维数据，其计算代价很高。另外不一致度的度量倾向于选择值变化较大的特征，如病人的身份证号码和病症无关，但不一致度评估则很可能会将身份证号码确认为相关特征。

·基于信息熵的评估标准 信息论中的熵是信息量的评估标准，信息熵的一个直观解释是描述信息所需要的二进制位数。对于离散的随机变量 x，Shannon 熵定义为：

$$H(x)=-\sum_{j=0}^{N-1}p(x_j)\log_2 p(x_j)$$

其中变量 x 的取值空间为$\{x_0,x_1,x_2,\cdots x_{N-1}\}$。基于信息熵有代表性的评估标准有：信息增益(Information Gain)、最小描述长度(Minimum description length)、互信息(Mutual Information)和关联度(Correlation)等。

信息增益是 Quinlan 在决策树算法中用来选择特征作为树分支结点时采用的评价标准[11]。其思想是考察类别的先验熵和已知一个属性的值 V 后的后验熵之间的差，并以之表示该属性所提供的类别可分性的信息。互信息在特征选择中常用来评价特征和目标属性间的紧密程度[37]。假定有两个事件，a 与 b，它们发生的机率分别是 $p(a)$与 $p(b)$，同时发生的机率是 $p(a,b)$。这两个事件的互信息定义为：$MI=\log_2[p(a,b)/(p(a)p(b))]$。MDL[31]评估标准是基于信息理论中的 Occam 的剃刀准则(Entities should not be multiplied beyond necessity)提出的，该标准争取用最小的编码长度来对所有的信

息进行编码，编码的长度由模型的编码长度和给定模型后数据的编码长度两部分组成，而具体模型或者给定模型后数据的编码长度则由相关信息熵表示。MDL 评估标准只适用于离散型特征，Domingos 指出对 Occam 剃刀准则的理解在理论上和实际应用中都不是绝对正确的[38]。MDL 评估标准和目标函数之间的关系不是直观的，其性能需要进一步进行实验验证和理论分析。信息增益、最小描述长度、Mutual Information，以及类似的评价标准如 Gini Index[12]以及 Lee 提出的模糊熵度量[33]都假定特征间彼此独立，因此不适用于特征相关性较强的领域，且只适用于离散型特征，其优点是实现简单，适用于大规模数据集，具有较强的理论基础。

Hall 提出的 Correlation 评估给出了一种既考虑了特征和目标函数的相关度，也考虑特征之间相关度的特征子集评估标准[39]。其准则为：好的特征子集含有的特征应该和目标函数有很高的相关度，但是和集合内其余特征不相关。但出于计算量限制，只能考虑两两特征之间的相关度，更复杂的特征之间的相互作用则不考虑。

·ReliefF 评估　Kira 等1992年提出的 RELIEF[40]是公认的性能较好的特征评估方法，其特征评价借用了最近邻学习算法的思想，其理论基础是一个好的特征应该使最近邻的同类样本之间特征值相同或相近，而使最近邻的不同类样本之间值不同或者差别很大。据此赋予每个特征相应权值进行特征排序，通过设定特征权阈值或者特征子集的数目即可进行相应的特征选择。Kononenko 对该方法进行了扩展[41]，使得 RELIEF 方法可以处理回归问题以及样本类别多于两类的情况。Marko 和 Kononenko 又从概率统计的角度对 RELIEF 方法进行了更深入的理论研究[42]，指出当样本无限多时，特征的权值满足如下逼近概率：W[特征 i]$=P$(特征 i 取值不同|不同类的最近邻样本)$-P$(特征 i 取值不同|同类的最近邻样本)，其中 W[特征 i]表示特征 i 的权值，P(特征 i 取值不同|不同类的最近邻样本)表示不同类的最近邻样本之间的特征 i 取值不同的概率。

·和学习算法性能相关的评估标准　又分为直接和间接利用学习算法性能作为特征评估标准两类。后者例如基于支持向量机的算法，提出的最大正边缘距离，最大负边缘距离等特征评估标准。还有将学习算法性能和特征子集大小综合考虑的特征评估标准[8,27,34,35]。

4　特征选择研究的最新现状，存在的不足和发展方向

近几年来，特征选择研究呈现出多样化和综合性的趋势。各种新搜索算法和评估标准都应用于特征选择，如粗糙集算法用于特征选择，神经网络剪枝法进行特征选择[43]，支持向量机的评估标准用于特征选择[34,35]，特征集的模糊熵评价等[33]。并且除监督式学习的特征选择研究外，也开展了关于非监督式学习的特征选择研究[44]。另外出现了关于特征选择的算法融合性的研究，如关于过滤式方法和 Wrapper 方法结合的研究[4,7]。

虽然特征选择领域的研究近年来取得了很大进展，但由于特征选择问题和实际对象的复杂性，还需要进行大量深入的研究，目前研究热点和需要解决的问题包括以下方面。

4.1　使用具有代表性的研究数据

目前，大部分关于特征选择的研究采用 UCI 机器学习数据库做为研究平台，而 Langley 等的实验表明 UCI 的大多数数据的特征数目较少[2]，且不相关特征很少，因为 UCI 数据多是由特定领域内的专家提供的，剔除了数据库中不相关的特征，而且数据量一般也不是很大。为研究特征选择算法在大规模数据集上的性能，可以选择特征数目较多、数据量较大的实际研究对象作为实验平台。另外，人工数据也可以为研究者提供一个好的实验平台。使用人工数据，研究者可以根据需要指定参数，而专门研究具体因素对特征选择算法的影响。

4.2　研究过滤式和 Wrapper 相结合的特征选择算法

过滤式方法快但准确率低，Wrapper 方法慢但准确率高。因此将过滤式方法和 Wrapper 方法结合起来以得到综合两者优点的特征选择方法是一个很有潜力的研究方向。1999年 Huang Yuan 等最早提出一种过滤式和 Wrapper 方法结合的方法[7]，其后越来越多的学者对该领域进行了研究[4,41]。但是目前关于此方面研究尚处于起步阶段，还需要借鉴样本选择以及集成学习等思想进行更系统的分析和研究。

4.3　研究特征选择和学习算法之间的关系

过滤式特征选择算法认为特征和后边的学习算法是独立的，即一个好的特征子集在任何学习算法上都应该取得较好的效果；Wrapper 算法认为特征选择不仅和数据集以及目标函数有关，还和后边的学习算法的性能密切相关，因此要用学习算法本身的性能作为选择特征子集的评估标准。目前对特征选择和学习算法之间的关系仍无定论，Kohavi 用几个实例证明过滤式算法的不足[8]，但是 Sanmay 认为 Kohavi 所举的都是很特殊的假想例子[45]，在实际情况中一个好的特征子集能使任何学习算法得到好的结果。进一步深入研究特征选择和学习算法之间的关系对于指导特征选择算法的构造有重要意义。

总结　本文对机器学习中特征选择算法和学习算法结合的三种方式，特征选择算法的四个要素进行了全面论述。介绍了特征选择的最新发展动态，并就特征选择研究的不足和可能的解决方案进行了讨论。

参 考 文 献

1　Langley P. Selection of relevant features in machine learning. In: Proc. AAAI Fall Symposium on Relevance, 1994. 140～144

2　Langley P, Iba W. Average-case analysis of a nearest neighbour algorithm. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1993, 2: 889～894

3　Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. IEEE transactions on pattern analysis and machine intelligence, 1997, 19(2): 153～158

4　Xing E, Jordan M, Karp R. Feature selection for high-dimensional genomic microarray data. Intl. conf. on Machine Learning, 2001. 601～608

5　章新. 一种特征选择的动态规划方法. 自动化学报, 1998, 25(4): 675～680

6　张鸿宾, 孙广煜. Tabu 搜索在特征选择中的应用. 自动化学报, 1999, 25(4): 457～466

7　Yuan H, et al. A two-phase feature selection methods using both filter and wrapper. In: IEEE SMC '99 Conf. Proc. 1999, 2: 132～136

8　Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence journal, special issue on relevance, 1997, 97(1－2): 273～324

9　Blum A L. Learning Boolean Functions in an Infinite Attribute Space. Machine Learning, 1992, 9(4): 373～356

10　Quinlan J R. Learning efficient classification procedures and their application to chess end games. Machine Learning: An artificial intelligence approach, San Francisco, CA: Morgan Kaufmann, 1983. 463～482

11　Quinlan J R. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann, 1993

12　Breiman L, Friedman J H, et al. Classification and Regression

Trees. Wadsforth International Group,1984
13 Cover T M. The best two independent measurements are not the two best. IEEE Trans. Syst. Man Cybern. ,1974,4(2):116～117
14 John G,Kohavi R,Pfleger K. Irrelevant features and the subset selection problem. The Eleventh International Conference on Machine Learning,1994. 121～129
15 Aha D W,Bankert R L. Feature selection for case-based classification of cloud types. In:working notes of the AAAI94 workshop on case-based reasoning,1994. 106～112
16 Provan G M,Singh M. Learning bayesian networks using feature selection. In: Proc. 5th Intern. Workshop on AI and Statistics, 1995. 450～456
17 Inza I,Larraaga P,Sierra B. Feature subset selection by Bayesian networks based on optimization. Artificial Intelligence,2001,123 (1-2):157～184
18 Caruana R A,Freitag D. Greedy attribute selection. The Eleventh intl. conf. on machine learning,1994. 28～36
19 Moore A W,Lee M S. Efficient algorithms for minimizing cross validation error. In: The Eleventh Intl. Conf. on Machine Learning,1994. 190～198
20 Davies S,Russl S. Np-completeness of searches for smallest possible feature sets. In:Proc. of the AAAI Fall 94 Symposium on Relevance,1994. 37～39
21 Liu H,Motoda H. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers,1998
22 Natendra P M,Fakunaga K. A branch and bound algorithm for feature subset selection. IEEE Trans. Comput. ,1977. 917～922
23 Pudil P,Novovicova J,Kittler J. Floating search methods in feature selection. Pattern Recognition Letters, 1994,15(11):1119～1125
24 Somol P,Pudil P,Novovicova J,Pacik P. Adaptive floating search methods in feature selection. Pattern Recogniton Letters,1999,20 (11－13):1157～1163
25 Yang J,Vasant H. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems,1998,13:44～49
26 Casillas J,Cordon O,et al. Genetic feature selection in a fuzzy rule-based classification system learning process for high dimensional problems. Information Sciences, 2001,136(1－4):135～157
27 Kudo M,Jack S. Comparison of algorithms that select features for pattern classifiers. Pattern Recogniton,2000,33:25～41
28 John G,Kohavi R,Pfleger K. Irrelevant features and the subset selection problem. In: The Eleventh Intl. Conf. on Machine Learning,1994. 121～129
29 Blum A L,Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997,97(2):245～271
30 Caruana R A,Freitag D. Ho v useful is relevance? Working notes of the AAAI Fall Symposium on Relevance. New Orleans,LA: AAAI Press,25～29
31 Pfahringer B. Compression-based feature subset selection. In:IJCAI-95 Workshop on data Engineering for Inductive Learning, 1995. 101～106
32 Liu H,Motoda H,Dash M. A monotonic measure for optimal feature selection. In: 10th European Conf. on Machine Learning, 1998. 101～106
33 Lee H M,Chen C M,et al. An efficient fuzzy classifier with feature selection based on fuzzy entropy. IEEE Trans. on systems and cybernetics-Part B: Cybernetics,2001,31(3):26～432
34 Weston J,Mukherjee S,et al. Feature selection for SVMS. In:Advances in Neural Information Processing Systems,2000,13:668～674
35 范劲松,方廷建.特征选择和提取要素的分析及其评价.计算机工程与应用,2001(13):95～99
36 Almullim H,Dietterich T G. Learning with many irrelevant features. In:Proc. Ninth National Conf. on Artificial Intelligence, 1991. 547～552
37 Hamming R W. Coding and information theory. Englewood Cliffs,NJ: Prentice-Hall,1986
38 Domingos P. The role of Occam's razor in knowledge discovery. Data Mining and knowledge Discovery,1999,3(4):409～425
39 Hall M A. Correlation-based feature selection for discrete and numeric class machine learning. In: The Seventeenth Intl. Conf. on Machine Learning,2000. 359～366
40 Kira K,Rendell L A. The feature selection problem: traditional methods and a new algorithm. In:Proc. of the Ninth National conf. on Artificial Intelligence,1992. 129～134
41 Kononenko I. Estimation attributes: analysis and extensions of RELIEF. In:Proc. of the 1994 European Conf. on Machine Learning,1994. 171～182
42 Marko R S,Kononenko I. Comprehensible interpretation of relief estimates. In:The Eighth Intl. Conf. on Machine Learning,2001. 443～440
43 Moore A W,Lee M S. Efficient algorithms for minimizing cross validation error. In:The Eleventh Intl. Conf. on Machine Learning,1994. 190～198
44 Dash M,Liu H. Feature selection for clustering. In:4th Pacific-Asia Conf on Knowledge Discovery and Data Mining,2000. 110～121
45 Sanmay D. Filters,Wrappers and a boosting-based hybrid for feature selection. In:Intl. conf. on Machine Learning,2001. 74～81

(上接第179页)

13 MaTait K. Translation Pattern Extraction and Recombination for EBMT:[PH. D thesis]. UMIST,2001
14 Frederking, Nirenburg. Three Heads are Better than One. In: proc. of ANLP-94,Germany,1994
15 Brant T. TNT - A Statistical Part-Of-Speech Tagger. In:proc. of the 6th Applied Natural Language Processing Conf. Seattle, Washington,USA,2000
16 Brill E. A Corpus-based Approach to Language Learning:[PH. D Thesis]. UPENN,1993
17 Daelemans W,Zavrel J,Berck P,et al. MBT: A Memory-Based Part of Speech Tagger Generator. In:Proc. of the Workshop on Very Large Corpora,Copenhagen,Denmark,1993
18 Ratnaparkhi A. A Maximum Entropy Model for Part-Of-Speech Tagging. In:proc. of EMNLP-96,Philadelphia,PA. 1996
19 Ng,Tou H,Lee H B. Integrating multiple knowledge sources to disambiguate word sense: An Example-Based Approach. In: the Proc. of the 34th ACL,1996. 40～47
20 Magerman D M. Statistical Decision-Tree Model s for Parsing. In: the Proc. of the 33rd ACL. Cambridge,MA,1995. 26～30
21 Ratnaparkhi A. A Linear Observed Time Statistical Parser Based on Maximum Entropy Models,In EMNLP2,Providence,RI. ,1997
22 Tenni J,Lehtola A,et al. Machine Learning of Language Translation Rules. In:IEEE Intl. Conf. on Systems,Man,and Cybernetics Tokyo International Forum,Tokyo,JAPAN,1999
23 Kishore,Salim,Todd,et al. Bleu: a Method for Automatic Evaluation of Machine Translation:[IBM Research Report]. 2001
24 utomatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics:[NIST Report]. 2001
25 Nomoto T. Predictive Models of Performance in Mulit-Engine Machine Translation,MT Summit IX,2003
26 Chen, Stanley F, Goodman J. An empirical study of smoothing techniques for language modeling:[Technical Report TR-10-98]. Center for Research in Computer Technology,Harvard University,1998
27 Charniak E,Knight K,Yamada K. Syntax-based Language Model for Statistical Machine Translation,MT Summit IX,2003
28 Brown R, Frederking. Applying statistical English Language Modelling to symbolic machine translation. TMI' 95, Leuven, Belgium,1995. 221～239