CN43-1258/TP
ISSN 1007-130X
计算机工程与科学
COMPUTER ENGINEERING & SCIENCE
2005 年第 27 卷第 12 期
Vol. 27, No. 12, 2005

文章编号:1007-130X(2005)12-0068-04

特征选择方法综述*
A Survey of Feature Selection

王 娟[1],慈林林[2],姚康泽[2]
WANG Juan[1], CI Lin-lin[2], YAO Kang-ze[2]
(1. 北京理工大学信息科学技术学院,北京 100081;2. 信息高技术研究所,北京 100085)
(1. School of Information Science and Technology, Beijing Institute of Techology, Beijing 10081;
2. Information High-Technology, Institute, Beijing 10085, China)

摘 要:本文总结并提出了较为完备的特征提取定义。根据特征子集形成过程将特征选择分为穷举式、启发式和随机式三类;根据特征评价标准将特征选择分为距离测度、信息测度、相关性测度、一致性测度和分类器错误率五类。通过分析特征选择的影响因素,提出了选择特征、选择方法应该遵循的原则。

Abstract: This paper analyzes and summarizes the previous definition of feature selection, and then introduces a self-contained definition. It divides feature selection into three classes according to the selecting strategy, and categorizes the methods into five styles by the evaluation function. Through analyzing the infection factors in the feature selection technology, this paper introduces some principles to pave the way for practitioners who search for suitable features to solve real-world applications.

关键词:特征选择;模式识别;特征评价
Key words: feature selection; recognition; feature evaluation
中图分类号:TP391.4 **文献标识码**:A

1 引言

特征选择在模式识别领域中扮演着一个极其重要的角色。一方面,在样本有限的情况下,用大量特征来设计分类器无论是从计算开销还是从分类器性能来看都不合时宜;另一方面,特征和分类器性能之间并不存在线性关系,当特征数量超过一定限度时,会导致分类器性能变坏。因此,进行正确有效的特征选择成为模式识别中必须要解决的问题,在海量数据条件下尤为重要。目前,国外有不少学者对此进行了研究,提出了许多方法[1~12];国内这方面研究还不很充分,多数情况下仍采用实验比较来选取特征。

本文主要目的是通过对各种特征选择方法的分类来给出解决特征选择问题的一般性原则。

2 特征获取过程

经典特征选择定义为从 N 个特征集合中选出 M 个特征的子集,并满足条件 $M \leqslant N$[1]。它包括特征提取和特征选择两个方面:特征提取广义上指的是一种变换,将处于高维空间的样本通过映射或变换的方式转换到低维空间,达到降维的目的;特征选择指从一组特征中去除冗余或不相关的特征来降维。二者常联合使用,如先通过变换将高维特征空间映射到低维特征空间,然后再去除冗余的和不相关的特征来进一步降低维数。本文用"特征获取"来统称特征选择和提取。

至今为止,有很多学者从不同角度出发对特征获取进行过定义:Kira[2]定义理想情况下的特征获取为寻找必要的、足以识别目标的最小尺寸特征子集;John[3]从提高预测精度角度定义特征获取为选择特征子集来增加分类精度,或者在不降低分类器精度的条件下降低特征集维数的过程;Koller[4]从类分布的角度定义特征获取为: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[5]给出的定义是选择尽量小尺寸的特征子集,并满足不显著降低分类精度和不显著改变类分

* 收稿日期:2004-03-16;修订日期:2004-05-28
基金项目:武器装备预先研究课题(403010503)
作者简介:王娟(1977-),女,湖南湘阴人,博士生,研究方向为图像处理、SAR 信息处理和模式识别。
通讯地址:100085 北京市海淀区清河上园 4-7-703 信箱;Tel:13661118469;E-mail:wangjuan@china.com.cn
Address: Mail Box 4-7-703, Qingshangyuan Garden, Qinghe Section, Haidian District, Beijing 100085, P. R. China

布两个条件。上述各种定义出发点不同，各有侧重点，但均未对特征子集的稳定性加以考虑。文献[6]证明了不同的分类器适应的特征组合和数目是不同的，即便是一个分类器获得最优结果的特征子集，不一定适用于其他分类器。根据这一现象，笔者认为特征获取的定义除了考虑对分类结果等的影响外，特征自身稳定性也是一个应该注意的因素，因此定义特征获取为获得尽可能小的特征子集的过程，并满足不显著降低分类精度、不影响类分布以及特征子集应具有稳定、适应性强的特点。

3 特征获取方法分类

特征获取需要解决两个问题，一是确定选择算法，在允许的时间内，以可以忍受的代价找出最小的、最能描述类别的特征组合；二是确定评价标准，衡量特征组合是否最优，得到特征获取操作的停止条件。因此，一般分两步进行特征获取，先产生特征子集，然后对子集进行评价，如果满足停止条件，则操作完毕，否则重复前述两步直到条件满足为止。下面将从这两方面对特征获取进行分类。

3.1 按照特征子集形成方式分类

按照特征子集的形成方式，特征获取方法可分为穷举法(Exhaustion)、启发法(Heuristic)和随机法(Random)三类[5]，如图1所示。

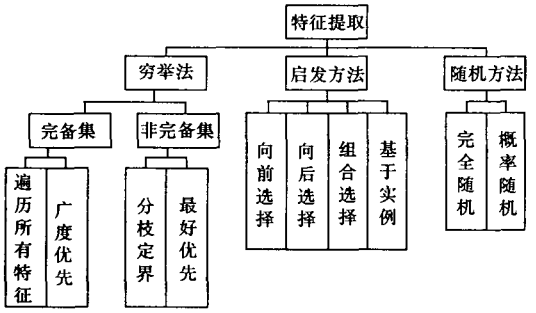

图1 按选择算法分类

穷举法指遍历特征空间中所有特征的组合，选取最优特征组合子集的方法。假设特征个数为 N 时，计算复杂度为 $O(2^N)$。常用的方法有回溯方法及其变体等[13]。其优点在于一定能得到最优子集，但实际情况下由于特征空间过于庞大，时间耗费和计算复杂度太大，导致实用性不强。

启发式方法为一种近似算法，具有很强的主观倾向。实际应用中通过采用期望的人工机器调度规则，重复迭代产生递增的特征子集。特征个数为 N 时，复杂度一般小于或者等于 $O(N^2)$。这种方法实现过程比较简单而且快速，在实际中应用非常广泛，如向前(向后)选择、决策树法[14]、Relief方法[2]及其变体[15]等。但是，不能保证结果最优，一般能够获得近似于最优解的解。

随机方法是一种相对较新的方法，细分为完全随机方法和概率随机方法两种。完全随机方法是指"纯"随机产生子集，概率随机是指子集的产生依照给定的概率进行。虽然计算复杂度仍为 $O(2^N)$，但通过设置最大迭代次数可以限制复杂度小于 $O(2^N)$。常用的方法有LVF(Las Vegas Filter，简称LVF)[16]、遗传算法[17]、模拟退火算法及其变体等。这类方法需要进行参数设置，并且参数值决定是否能得到最优解。如何有效地设置这些参数是一个值得研究的问题。

总的说来，上述三类中只有穷举法能保证最优，但耗时并且计算复杂度很高，后两者以性能为代价换取简单、快速的实现，但不能保证最优。实际应用中为了折衷性能和代价之间的矛盾，常结合几种方法，如文献[18]中采用三步法：首先使用Relief算法去除无关的特征，其次采用 k 均值法去除冗余特征，然后进行标准的组合特征方法，取得了较好的效果。这也是进一步研究的方向。

3.2 按照特征评价标准分类

特征选择可以看作一个优化问题，其关键是建立一种评价标准来区分哪些特征组合有助于分类，哪些特征组合存在冗余性、部分或者完全无关。不同的评价函数可能会给出不同的结果。根据评价函数与分类器的关系，特征选择方法分成筛选器[16,19]和封装器[20,21]两种。其中，筛选器的评价函数与分类器无关，而封装器采用分类器的错误概率作为评价函数。其中，筛选器的评价函数又可以细分为距离测度、信息测度、相关性测度和一致性测度。

特征获取的最终目的在于使分类器的错误概率最小，因此最直观的方式是采用分类器错误概率作为评价标准，即选择使分类器的错误概率最小的特征或者特征组合。Solberg[22]等人对SAR图像多纹理特征分类、Sylbie[23]对机载多光谱和多频率SAR数据进行无监督分类，都是通过分类结果比较来选择特征。但是，这种方法计算量太大，实用性差，即使在类条件分布密度已知的情况下计算分类器的错误概率都十分复杂，而实际中往往类条件分布密度未知，更加难以计算。因此，基于评价函数进行特征选择更为常用。

距离测度是利用距离来度量样本之间相似度的一种方式。分布于不同区域的样本，样本之间距离越小越相似，样本之间距离越大，其可分性就越大。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、非线性测量等，其中欧氏距离可以看作是2阶Minkowski距离。直接从样本间的距离计算获取的距离判据虽然计算方便，直观概念清楚，但没有考虑各类的概率分布，不能确切表明各类交叠的情况。因此，概率距离测度作为一种扩展被提出。常用的概率距离测度有Bhattacharyya距离、散度、Chernoff概率距离以及Mahalanobis距离等。

信息测度是为了衡量后验概率分布的集中程度所规定的一个定量指标。从特征获取的角度来看，利用具有最小不确定性的那些特征来分类是最有利的，因此引入信息领域中作为不确定性量度的熵函数作为评价测度[24]。常用的熵函数有Shannon熵、Renyi熵和条件熵等。

除了最常用的距离测度和信息测度以外，还有两种较为常见的测度：相关性测度和一致性测度。相关性测度包括两个方面的内容，既可以利用相关系数，找出特征和类之间存在的相互关系；又可以利用特征之间的依赖关系，来表示特征的冗余性，文献[25]对此有详细论述。一致性测度发展较晚，它和训练数据集关系密切，并且需要设定参数，最后得到的结果为满足给定参数的最小尺寸特征子集。文献[16]利用不一致率作为阈值来进行特征选择。

尽管上述多种评价函数具有各自的特点，如表1所示，

但其中一部分评价函数可以看作是另一部分的数学转换，两者之间存在等价关系。值得注意的是，上述评价函数多数情况下和分类错误概率没有直接关系。因此，根据获取结果所设计的分类器的错误概率未必是最小的。但是，从简化和降低计算的观点来看，应用其中的一些评价函数进行特征获取往往是唯一可行的办法。

表 1 评价函数性能指标

评价函数	泛化能力	时间复杂性	分类精度
距离测度	好	低	—
信息测度	好	低	—
相关性测度	好	低	—
一致性测度	好	中等	—
分类错误率	差	高	很好

4 特征获取方法的选择原则

理想的特征获取是严格筛选尽量少的、最佳的、最有影响的特征集合，实现最简单方便的分类。良好的特征集合应具有可辨别性好、可靠性高、独立性强、稳定性高和数量少等特点。目前还没有文献给出一种简单而实用的特征获取方法来解决特征选择问题。下文将从影响特征选择的因素和选取原则两个方面来讨论此问题。

4.1 影响特征选择方法的因素

影响特征选取方法的因素主要有数据类型、问题规模、样本数量等[5]，如图 2 所示。此外，噪声引入的一些虚假或冗余特征，也会导致数据矛盾甚至错误，极大影响分类结果。因此，对噪声的容忍能力成为确定特征选择方法的考虑因素。

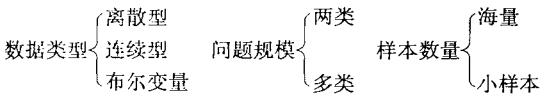

图 2 影响特征选择的因素

4.2 选取原则

基于上述影响因素，确定合适的特征获取方法应该遵循以下一些原则：

(1)处理数据类型的能力。判断是否支持离散数据、连续数据或布尔类型数据。各种特征选择方法有其处理数据类型的范围，如分枝定界法不支持布尔类型，Koller-Sahami's[4]不支持连续类型等。

(2)处理问题规模的能力。判断是否能够处理两类问题或者多类问题，如 Relief[2]不支持多类问题等。一般情况下，可以先将多类问题划分为若干个两类问题，然后利用两类问题的选择方法进行处理来扩展处理能力。

(3)处理样本数量的能力。判断是否能够处理小样本数据集或海量数据。有文献表明，特征选取方法对于特征集的大小有限制，如 SFBS 不能适应特征个数多于 110 的特征集[26]。

(4)对噪声的容忍能力。实际问题情况十分复杂，噪声分布各不相同，有强有弱。一般是抗噪性越强，获取特征的性能也就越好。

(5)无噪声情况下，产生稳定的、最优特征子集的能力。所谓最优特征子集的产生能力，除了直接由结果最优来决定外，还需要考虑代价因素。只要在允许的代价下能够获取满足要求的结果，就可以视为最优。因此，关于最优特征子集的衡量和实际参数有关。

各类评价函数的性能指标的影响如表 1 所示。对实时性较强的处理操作，最好不要采用分类错误概率评价。

5 结束语

特征获取是模式识别过程中必不可少的一环，受到广泛的关注。本文结合文献中出现的一些特征选择方法，按照选择算法将特征获取方法分为穷举法、启发法和随机方法三种；按照特征评价函数将特征获取分为距离测度、信息测度、相关性测度、一致性测度和分类器错误率五类；并且，通过分析影响特征获取算法的因素，提出选取算法应该遵循的原则。尽管已经有很多进行特征获取的方法，但针对解决实际问题的研究还很不充分。如何组合利用现有的方法，以及针对特定问题提出新的方法，是目前特征获取方法研究的发展方向。

参考文献：

[1] 边肇祺，张学工．模式识别．第 2 版[M]．北京：清华大学出版社，2000.

[2] K Kira, L A Rendell. The Feature Selection Problem: Traditional Methods and a New Algorithm[A]. Proc of 9th National Conf on AI[C]. 1992. 129-134.

[3] G H John, R Kohavi, K Pfleger. Irrelevant Features and the Subset Selection Problem[A]. Proc of the 11th Int'l Conf on Machine Learning[C]. 1994. 121-129.

[4] D Koller, M Sahami. Toward Optimal Feature Selection[A]. Proc of Int'l Conf on Machine Learning[C]. 1996. 284-292.

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

[6] Reinhold Huber, Luciano V Dutra. Feature Selection for ERS-1/2 InSAR Classification: High Dimensionality Case[A]. Proc of Int'l Geoscience and Remote Sensing Symp Proceedings. Vol 3[C]. 1998. 1605 - 1607.

[7] Y Yamagata, H Oguma. Bayesian Feature Selection for Classifying Multi-Temporal SAR and TM Data[A]. Proc of Int'l Geoscience and Remote Sensing Symp. Vol 2[C]. 1997. 978 - 980.

[8] A L Blum, P Langley. Selection of Relevant Feature and Examples in Machine Learning[J]. Artificial Intelligence, 1997, 97: 245-271.

[9] M Scherf, W Brauer. Feature Selection by Means of a Feature Weighting Approach[Z]. Technical University Munchen, 1997.

[10] B Chakraborty. Genetic Algorithm with Fuzzy Fitness Function for Feature Selection[A]. Proc of the 2002 IEEE Int'l Symp on Industrial Electronics. Vol 1[C]. 2002. 315-319.

[11] S B Serpico, L Bruzzone. A New Search Algorithm for Feature Selection in Hyper Spectral Remote Sensing Images[J]. IEEE Trans on Geoscience and Remote Sensing, 2001, 39(7): 1360-1367.

[12] 谭湘莹，于秀兰，钱国惠．一种大小窗口结合的 SAR 图像纹理特征分类方法[J]．系统工程与电子技术，2000，22(4)：15-17.

[13] L Xu, P Yan, T Chang. Best First Strategy for Feature Selection[A]. Proc of 9th Int'l Conf on Pattern Recognition[C]. 1988. 706-708.

[14] C Cardie. Using Decision Trees to Improve Case-Based Learning[A]. Proc of 10th In'l Conf on Machine Learning [C]. 1993. 25-32.

[15] I Kononenko. Estimating Attributes: Analysis and Extension of Relief[A]. Proc of European Conf on Machine Learning [C]. 1994. 171-182

[16] H Liu, R Setiono. A Probabilistic Approach to Feature Selection: A filter Solution[A]. Proc of Int'l Conf on Machine Learning[C]. 1996. 319-327.

[17] B Chakraborty. Genetic Algorithm with Fuzzy Fitness Function for Feature Selection[A]. Proc of the 2002 IEEE International Symp on Industrial Electronics. Vol 1[C]. 2002. 315 - 319.

[18] Jos Bins, Bruce A Draper. Feature Selection from Huge Feature Sets[A]. Proc of the 8th IEEE Conf on Computer Vision and Pattern Recognition. Vol 2[C]. 2001. 159-165.

[19] Sanmay Das. Filters, Wrappers and a Boosting Based Hybrid for Feature Selection[A]. Proc of the 8th Int'l Conf on Machine Learning[C]. 2001. 74-81.

[20] Huang Yuan, Shian-Shyong Tseng, Wu Gangshan, et al. A Two-Phase Feature Selection Method Using Both Filter and Wrapper[A]. Proc of 1999 IEEE Inter'l Conf on Systems, Man, and Cybernetics. Vol 2[C]. 1999. 132 - 136.

[21] R Kohavi, G H John. Wrappers for Feature Subset Selection [J]. Artificial Intelligence Journal, 1997, 97(1-2): 273-324.

[22] Solberg A H S, Jain A K. Texture Fusion and Feature Selection Applied to SAR Imagery[J]. IEEE Trans on Geoscience and Remote Sensing, 1997, 35(2): 475-479.

[23] Sylvie Le Hegarat-Mascle, Isabelle Bloch, et al. Application of Dempster-Shafer Evidence Theory to Unsupervised Classification in Multi-Source Remote Sensing[J]. IEEE Trans on Geoscience and Remote Sensing, 1997, 35(4): 1018-1031.

[24] Kari Torkkola. Nonlinear Feature Transforms Using Maximum Mutual Information[A]. Proc IJCNN[C]. 2001. 2756-2761.

[25] M Ben-Bassat. Pattern Recognition and Reduction of Dimensionality[A]. P R Krishnaiah, L N Kanal, ed. Handbook of Statistics[M]. 1982. 773-791.

[26] Bins, Bruce A. Draper Feature Selection from Huge Feature Sets[EB/OL] http://www.cs.colostate.edu/~draper/publications/bins_iccv01.pdf, 2003-12.

[27] 钱忠良，王文军. 不变矩目标特征描述误差分析和基于上层建筑不变矩的舰船识别[J]. 电子测量与仪器学报，1994，8(3)：23-31.

[28] 赵小杰，钟劲松，王宏琦. 合成孔径雷达图像的特征选择[J]. 遥感技术与应用，2001，16(3)：190-194.

(上接第 44 页)

表 1 执行数据划分算法之后进行联接查询的查询时间 s

节点数	查询时间	
	加强的混合范围划分方法	改进的混合范围划分方法
10	8.2	7.8
20	7.4	5.6
30	6.8	3.8

从表 1 的结果还可以看出：在节点数的增加对于查询性能提高的效果方面，IHRPS 要比 EHRPS 明显。对于 EHRPS，引入更多的节点，查询时间并不会有明显的缩短；而且还存在着查询负载过于集中的可能，极大地降低了系统的查询性能。而对于 IHRPS，随着节点数的增加，在保证各节点查询负载减少的同时，还能够保证节点间查询负载的平衡，提高查询的并行度；在最极端的情况下，查询所操作的几个关系的所有数据分块，其所在节点集没有交集，这样可以得到最高的并行度。所以，增加节点对于 IHRPS 的性能提高是十分明显的。

下面将对数据迁移算法进行测试。在利用数据划分算法进行初期的数据分置之后，反复执行联接查询 S_2，并且不断改变 S_2 的查询参数 Max_a、Min_a 和 Max_b、Min_b。这样做的目的是将改变各节点的热度，使得系统中节点之间的热度差别变大，直到大于 K_{ndef}。然后，执行数据迁移算法，并记录算法的执行时间，最后用联接查询 S_2 测试迁移后的性能。

结果如表 2 所示，可以看出：数据迁移算法能够比较好地解决数据库系统长期运行后各节点之间热度差别过大的问题；但是，数据迁移算法的代价较大，而且节点越多，这一代价越高。所以，对于数据迁移算法，应该选择一个系统负载较小的时间段来运行。

表 2 数据迁移算法性能测试 s

节点数	查询时间			迁移时间
	数据划分后	多次查询后	数据迁移后	
10	8.0	12.5	9.1	210.8
20	6.2	10.3	7.9	275.9
30	4.4	8.7	6.5	325.1

5 结束语

通过改进，并行实时数据库系统不仅达到了节点间数据存储量的平衡，还达到了运行时查询负载的动态平衡，提高了系统的性能。但是，即使经过了上述改进，以混合范围划分方法为基础的并行实时数据库系统数据划分方法还不能满足实际需求。对于并行实时数据库系统的数据分置方法的研究还需继续进行。

参考文献：

[1] S Ghandeharizadeh, J D Dewitt. Hybrid Range Partitioning Strategy: A New Declustering Strategy for Multiprocessor Database Machines[A]. Proc of the 16th VLDB Conf[C]. 1990. 481-492.

[2] Khanh Quoc Nguyen, T Thompson, G Bryan. An Enhanced Hybrid Range Partitioning Strategy for Parallel Database [A]. Proc 8th Int'l Workshop on Database and Expert Systems Applications[C]. 1997. 289-294.

[3] D Dewitt, S Ghandeharizadeh, D Schneider, et al. The Gamma Database Machine Project[J]. IEEE Trans on Knowledge and Data Engineering, 1990, 2(1): 44-62.

[4] George Copeland, William Alexander, Ellen Boughter, et al. Data Placement in Bubba[J]. ACM SIGMOD Record, 1998, 17(3): 99-108.