一种基于信息增益及遗传算法的特征选择算法*)

任江涛　孙婧昊　黄焕宇　印　鉴
(中山大学计算机科学系　广州 510275)

摘　要　特征选择是模式识别及数据挖掘等领域的重要问题之一。针对高维数据对象，特征选择一方面可以提高分类精度和效率，另一方面可以找出富含信息的特征子集。针对此问题，本文提出一种综合了 filter 模型及 wrapper 模型的特征选择方法，首先基于特征之间的信息增益进行特征分组及筛选，然后针对经过筛选而精简的特征子集采用遗传算法进行随机搜索，并采用感知器模型的分类错误率作为评价指标。实验结果表明，该算法可有效地找出具有较好的线性可分离性的特征子集，从而实现降维并提高分类精度。

关键词　特征选择，信息增益，遗传算法

Feature Selection Based on Information Gain and GA

REN Jiang-Tao　SUN Jing-Hao　HUANG Huan-Yu　YIN Jian
(Department of Computer Science, Zhongshan University, Guangzhou 510275)

Abstract　Feature selection is one of the important problems in the pattern recognition and data mining areas. For high-dimensional data, feature selection not only can improve the accuracy and efficiency of classification, but also can discover informative feature subset. This paper proposes a new feature selection method combining filter and wrapper models, which first filters features by feature partition based on information gain, and realizes the near optimal feature subset search on the compact representative feature subset by genetic algorithm; and the feature subset is evaluated by the classification inaccuracy of the perceptron model. The experiments show that the proposed algorithm can find the feature subsets with good linear separability, which results in the low-dimensional data and the good classification accuracy.

Keywords　Feature selection, Information gain, GA

1　引言

特征选择是模式识别与数据挖掘领域的重要数据处理方法之一。随着模式识别与数据挖掘研究的深入，研究对象越来越复杂，对象的特征维数越来越高。大量高维数据对象的特征空间中含有许多冗余特征甚至噪声特征，这些特征一方面可能降低分类或聚类的精度，另一方面会大大增加学习及训练的时间及空间复杂度。因此，在面对高维数据进行分类或聚类时，通常需要运用特征选择算法找到具有较好可分性的特征子空间，从而实现降维，降低机器学习的时间及空间复杂度[1,2,8]。

根据是否依赖机器学习算法，特征选择算法可以分为两大类，一类为 wrapper 型算法，另一类为 filter 型算法。Filter 型特征选择算法独立于机器学习算法，具有计算代价小，效率高但降维效果一般等特点；而 wrapper 型特征选择算法则需要依赖某种或多种机器学习算法，具有计算代价大，效率低但降维效果好等特点。[1,2]

从优化的观点来看，特征选择问题实际上是一个组合优化问题。通常解决该问题有遍历搜索、随机搜索及启发式搜索等方法。遗传算法在组合优化问题中也有着广泛的应用，属于一种随机搜索方法。近年来，随着对特征选择方法研究的深入，基于遗传算法的特征选择问题也得到了许多研究及应用[4]。目前基于遗传算法的特征选择方法通常基于分类器进行特征子集的评估，依据分类精度给出个体的评价指标及适应度。

本研究融合特征选择算法的 filter 模型及 wrapper 模型，提出了一种基于信息增益及遗传算法的特征选择方法。首先基于特征之间的信息增益进行特征分组及筛选，然后针对经过筛选而精简的特征子集采用遗传算法进行随机搜索，并采用感知器模型的分类错误率作为评价指标。另外，在遗传算法编码方面没有采用传统的二进制直接编码方案，而是采用基于区间的二进制编码方案，一方面减小了编码长度、提高了时空效率，另一方面可对选择的特征个数进行灵活控制。

论文的第 2 部分首先简要介绍了相关工作及背景，第 3 部分对所提出的算法进行了描述，第 4 部分给出了实验研究结果，最后是对本文的总结。

2　基于信息增益的特征分组及筛选

特征分组是进行特征选择及降维的有效方法之一，其主要思想是基于特定的相似性度量，对特征进行分组，使得分在同一组的特征具有很强的相似性，而不同组的特征具有较大的差异，然后选出各组的代表特征作为精简后的特征子集，从而在一定程度上消除特征冗余，实现降维。在本研究中，采用信息增益作为特征之间的相似性度量，并采用一种基于密度的分组方法进行特征分组，实现特征的精简，下面首先给出信息增益的定义[5,6]。

*)本文研究得到国家自然科学基金资助(60573097)、广东省自然科学基金资助(05200302、04300462)。任江涛　博士，讲师。

令 X 为随机变量，则 X 的信息熵定义为：

$$H(X)=-\sum_i P(x_i)\log_2(P(x_i)) \quad (1)$$

通过观测随机变量 Y 随机变量 X 的信息熵变为：

$$H(X|Y)=-\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \quad (2)$$

其中 $P(x_i)$ 代表随机变量 X 的先验概率，$P(x_i|y_j)$ 代表观测到随机变量 Y 后随机变量 X 的后验概率。引入随机变量 Y 的信息后，随机变量 X 的信息熵 $H(X|Y)\leqslant H(X)$，即引入 Y 后，X 的不确定程度会变小或保持不变。若 Y 与 X 不相关，则 $H(X|Y)=H(X)$；若 Y 与 X 相关，则 $H(X|Y)<H(X)$，而差值 $H(X)-H(X|Y)$ 越大，Y 与 X 的相关性越强。因此，式(3)定义信息增益 $IG(X|Y)$ 为 $H(X)$ 与 $H(X|Y)$ 的差值，反映了 Y 与 X 的相关程度，$IG(X|Y)$ 越大，则变量 Y 与 X 的相关性越强。

$$IG(X|Y)=H(X)-H(X|Y) \quad (3)$$

而且，可以证明，信息增益具有对称性，即 $IG(X|Y)=IG(Y|X)$。另外，为了对信息增益进行归一化，可采用式(4)，同理有 $SU(X,Y)=SU(Y,X)$。

$$SU(X,Y)=2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right] \quad (4)$$

在相似度定义的基础上，就可以基于特征之间的相似度进行特征分组。本研究采用一种基于密度的分组方法，该方法的主要思路是首先针对每个特征分别统计与该特征相似度大于某个阈值的其它特征的个数，然后找出密度最大(即与该特征相似度大于指定阈值的其它特征数最大)的特征，将该特征及与其相似度大于指定阈值的其它特征归为一组(该特征即为此组特征的代表特征)；然后将该组特征从原特征集合中删除，继续上述过程，直至所有特征都被归到某一特定组为止；最后各组特征的代表特征即形成精简后的特征子集，具体算法流程如算法 1 所示。

算法 1：*PartitionFeatures*(F,δ)

输入：原始特征集合 F，阈值 δ

输出：精简后的特征子集 FS

步骤：

1)初始化特征子集 $FS=\{\}$，$FW=F$；

2)根据式(1)～(4)计算每个特征与其它特征的信息增益，形成特征相似度(信息增益)矩阵 SU；

3)针对每个特征 f_i 在特征集合 FW 中搜索与其信息增益大于阈值 δ 的其它特征，形成特征子集 F_i 如下：

$F_i=\{f_k|SU(f_i,f_k)\geqslant\delta, f_k\in FW, k\neq i\}, i=1,2,\cdots,|FW|$；

4)令 $S_i=|F_i|, i=1,2,\cdots,|FW|$；

5)令 $S_m=\mathrm{Max}(S_i), i=1,2,\cdots,|FW|$；

6)将特征 f_m 选入代表特征子集 FS，即 $FS=FS\cup\{f_m\}$；

7)从特征集合 FW 中剔除特征子集 F_m 及特征 f_m，即 $FW=FW-(F_m\cup\{f_m\})$；

8)重复步骤 3)～7)，直至 $FW=\{\}$ 时结束，输出精简的代表特征集合 FS。

3 基于遗传算法的特征选择

第 2 节的算法 1 通过消除特征之间的冗余性实现了原始特征集合的精简后，可以采用基于遗传算法的 Wrapper 型特征选择方法。下面从编码方案、适应度函数及算法流程等方面对该算法进行描述。

3.1 编码方案

编码问题的关键在于能代表所给特征集合的所有可能子集的解空间。常用的方法是采用直接二进制编码，即每一个二进制位对应特征集合中的一个特征，该位为 1 则表示对应的特征入选特征子集，而该位为 0 则表示对应的特征不在选出的特征子集中。在特征维数 d 相对较低时，该表示方法可得到较小的二进制串，提高计算效率。但在特征维数 d 特别高的情况下，该表示方法反而可能导致较长的串，从而降低了计算效率。例如，基因表达数据集 Colon Tumor 的维数为 2000，采用直接二进制的编码方法就需要长度为 2000 的二进制串。另外，直接的二进制表示方法不利于对选择出的特征个数进行限制，因此本研究采用基于区间的二进制编码方案，即用一个长度为 l 的二进制数表示所选择的特征在原特征集合中的序号。这样，如果指定要选择的特征个数 j，则这个二进制串长度为 $j*l$。当 $j\ll d$ 时，可得到较小的二进制串。例如，针对 2000 个特征，每个特征编号需要一个 11 位的二进制数串来表示，即 $l=11$，假设每次搜索 6 个特征的组合($j=6$)，那么整个编码二进制串的长度为 $j*l=6*11=66$，大大小于直接二进制编码的长度 2000，提高了空间及时间效率。同时，该编码方案可保证每次选择的特征个数可指定，从而实现了对特征子集大小的灵活控制。

3.2 适应度定义

在大多数基于遗传算法的 Wrapper 型特征选择方法中，采用某些分类器模型对所选择的特征集合进行评价，并利用得到的分类精度或分类错误率作为适应度函数。在本研究中，为搜索出线性可分性较好的特征子空间，采用感知器模型作为分类器模型，并采用分类错误率作为适应度，评价算法 *Evaluation* 的流程由算法 2 给出。

算法 2：*Evaluation*(D,F_s)

输入：数据集 D，特征子集 F_s

输出：特征子集评价值

步骤：

1)根据特征子集 F_s 从数据集 D 中选出一个降维后的数据集 D_F；

2)采用感知器算法对数据集 D_F 进行分类，统计分类错误率 err；

3)输出分类错误率 err 作为特征子集的评价指标，即适应度，算法结束。

4 算法流程

根据第 2、3 节的讨论，基于标准遗传算法框架，得到一种新的基于信息增益及遗传算法的特征选择方法(Feature Selection based on Information Gain and GA——FSIGGA)，算法具体描述如下。

算法 3：*FSIGGA*($D,F,\delta,fn,\mathrm{Max}I$)

输入：数据集 D，特征集合 F，阈值 δ，选择的特征数 fn，最大迭代次数 $\mathrm{Max}I$

输出：优化的特征子集

步骤：

(1)调用算法 *PartitionFeatures*(F,δ)，进行特征的过滤，形成精简特征子集 F_c；

(2)根据 3.1 节给出的编码方案，以及选择的特征数 fn，随机产生一组初始个体构成初始种群；

(3)根据编码方案，将个体的二进制表达转化为精简特征子集 F_c 中的特征编号，根据这些特征编号进行特征选择，形成特征子集 F_s；

(4)根据 3.2 节给出的适应度评价方法，调用函数 *Evaluation*(D,F_s)，计算个体适应度；

(5)判断是否达到最大迭代次数 Max$I$，若达到则输出当前的最优特征子集，否则执行以下步骤；

(6)根据适应度执行选择操作；

(7)执行交叉操作；

(8)执行变异操作；

(9)返回步骤(3)。

5 实验研究

为了评估上述 FSIGGA 算法的有效性，采用基因表达数据集 Colon Tumor 进行实验研究。Colon Tumor 数据集有 62 个样本，其中 40 个样本来自结肠癌患者的肿瘤组织，另外 20 个样本来自结肠癌患者的正常组织。数据集含有 2000 个基因，即特征数为 2000。

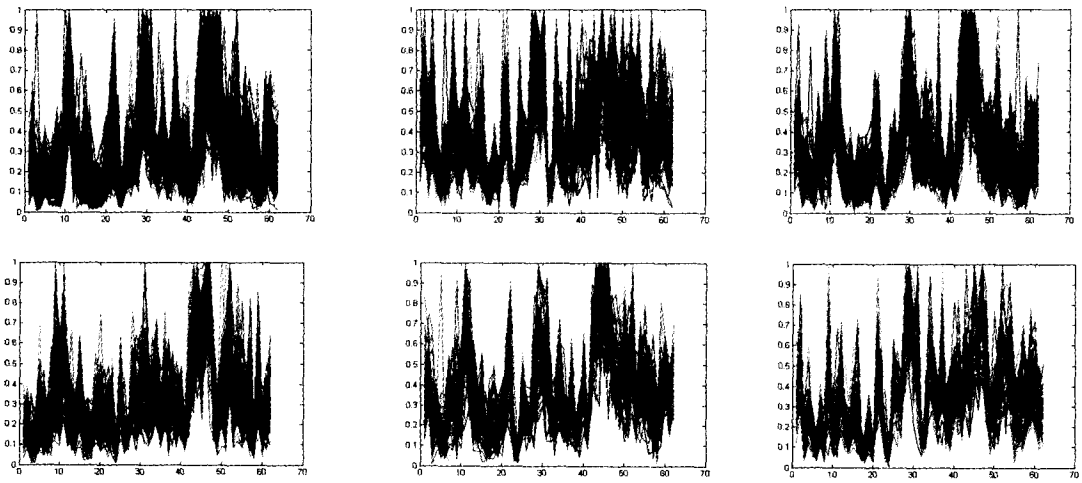

图 1　6 组具有强相关性的典型特征集合示例

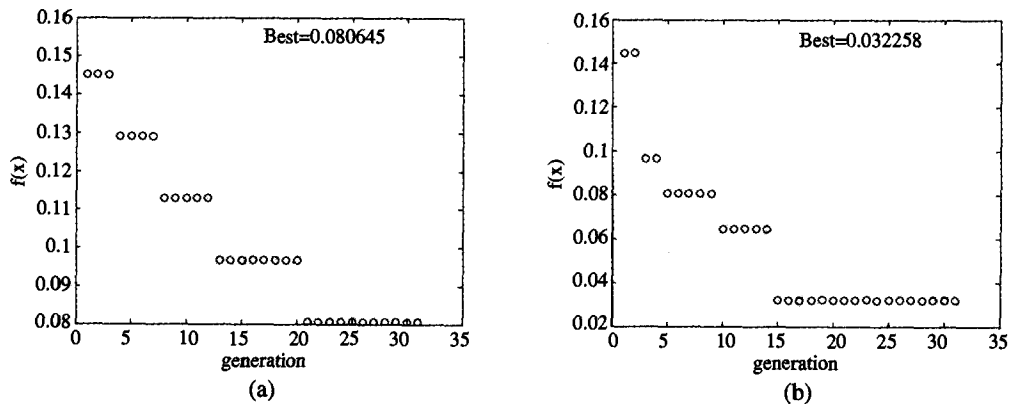

图 2　FSIGGA 算法针对 Colon Tumor 数据集的遗传算法迭代运行结果，特征维数分别为 2(a)及 3(b)

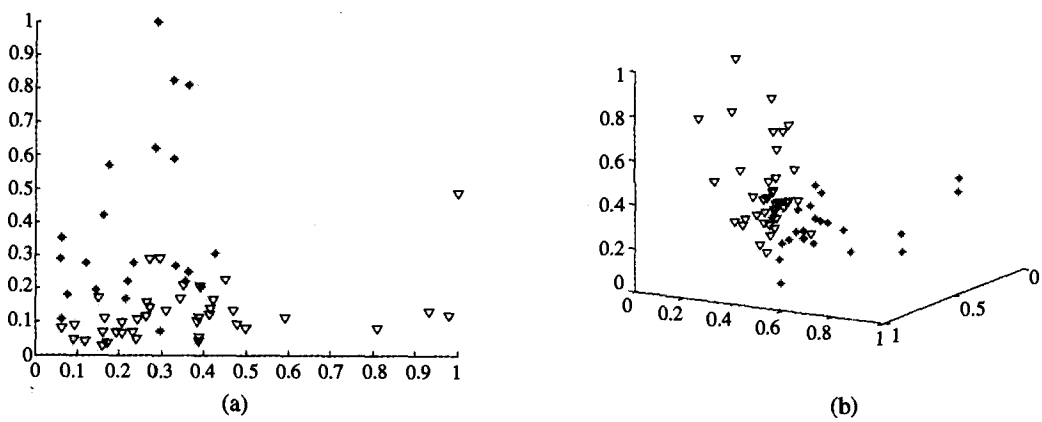

图 3　Colon Tumor 数据集在 FSIGGA 算法选出的 2 维(a)及 3 维(b)特征子空间的散点图

实验首先运用 PartitionFeatures 算法对特征进行分组，相似度阈值为 0.45，最后得到 592 组特征，图 1 的 6 个子图所示为其中 6 组典型特征集合的曲线，横坐标代表样本编号，纵

(下转第 251 页)

可能得到满足，不但与其领域相关，而且与该领域内其它涉众对其观点是否认同相关。

现有的多视点需求工程方法中关于不一致性处理的方法还很少有对认知属性进行研究的。但是我们相信，在需求工程中不一致性的问题是由于涉众对某一问题存在着不同的理解所导致的，也就是说这是一个与认知相关的问题。所以我们提出了基于问题域的多视点建模框架，并使用认知逻辑对该框架进行解释，从而提供了一种形式化的方法来发现和处理不一致性问题，希望这种方法可以提高需求规约说明书的质量，为开发出满足涉众需求的软件系统奠定基础。

本文的主要贡献在于提出了一个基于问题域的多视点建模框架，并使用认知逻辑对该框架进行解释。这不但符合软件开发活动中涉众处于不同领域的实际，同时也提供了一个形式化的方法来帮助解决多视点需求工程方法中必须要解决的不一致性问题。

在我们现有的工作中，还有许多问题需要在以后的工作中进一步去解决。譬如：知识应该是一个动态的过程。涉众在不同的时间对预期系统有着不同的看法。也就是说，不一致性的问题应该是随着时间的改变而改变，它具有动态性，而不能仅仅从静态的视角观察它。但是在我们这个框架和解释下还不能表示知识是与时间相关这一特性的。如果不能解决这一问题，就难以真正处理变化的需求以及做到需求的可追踪性。这是我们以后需要进一步研究的课题。

参 考 文 献

1 Finkelstein A, Gabbay D, Hunter A, et al. Inconsistency handling in multiperspective specifications. IEEE Trans. on Software Engineering, 1994, 20(8): 569～578

2 Zave P. Classification of Research Efforts in Requirements Engineering. ACM Computing Surveys, 1997, 29(4): 315～321

3 Balzer R. Tolerating inconsistency. In: Proceedings of the Fifth International Software Process Workshop (ISPW '89), Kennebunkport, Maine, USA. IEEE Computer Society 1989. 41～42

4 Ainsworth M, Cruickshank A H, Groves L J, Wallis P J L. Formal specification via viewpoints. In: Proceedings of the 13th New Zealand Computer Conference New Zealand Computer Society, Auckland, New Zealand, 1993. 218～237

5 Jackson D. Structuring Z specifications with views. ACM Transactions on Software Engineering and Methodology. , 1995, 4: 365～389

6 Boiten E, Derrick J, Bowman H, Steen M. Consistency and refinement for partial specification in Z. In: Proceedings of the Third International Symposium of Formal Methods Europe (FME'96): Industrial Benefit of Formal Methods, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1996, 1051: 287～306

7 Leduc G. On the role of implementation relations in the design of distributed systems using LOTOS, Ph. D. Thesis, University of Liège, Liège, Belgium, 1991

8 Khendek F, von Bochmann G. Merging specification behaviors, Technical Report 856, Departement d'informatique et de recherche operationnelle, Universite de Montreal, 1993

9 Ichikawa H, Yamanaka K, Kato J. Incremental specification in LOTOS, In: Proceedings of the IFIP WG6. 1 Tenth International Symposium on Protocol Specification, Testing and Verification X North-Holland Ottawa, Canada, 1990. 183～196

10 Bowman H, Derrick J, Linington P, Steen M W A. FDTs for ODP Computer Standards and Interfaces, 1995, 17: 457～479

11 Easterbrook S, Chechik M. A Framework for Multi-Valued Reasoning over Inconsistent Viewpoints. In: Proceedings of the 23rd International Conference on Software Engineering(ICSE'01)(Toronto, Ontario, Canada May 12-19, IEEE Computer Society, 2001. 411～ 420

12 Sabetzadeh M, Easterbrook S M. Analysis of Inconsistency in Graph-Based Viewpoints: A Category-Theoretic Approach. In : Proceedings of the 18th IEEE Int. Conf. on Automated Software Engineering(ASE 2003) (Montreal, Canda, Oct. 6-10), IEEE Computer Society, 2003. 12～21

13 Nuseibeh B, Kramer J, Hunter A. A framework for expressing the relationships between multiple views in requirements specification. IEEE Trans. on Software Engineering, 1994, 20(10): 760～773

14 Fagin R, Halpern J Y, Moses Y, Vardi M Y. Reasoning about Knowledge. The MIT Press: Cambridge, MA, 1995

15 Nuseibeh B, Finkelstein A, Kramer J. Viewpoints: Meaningful relationships are difficult! In: Proceedings of the 27th Int'l Conf. on Software Engineering. Oregon: IEEE Computer Press, 2003. 676～683

(上接第 195 页)

坐标代表特征(基因)在该样本上的取值(基因表达值)，被分至同一组的特征曲线在同一个图上，从图中可以看出，被分至同一组的特征均具有较强的相关性，因此可以通过保留每组代表特征并去除组内其余特征的方式消除特征冗余。图 2 给出采用 FSIGGA 算法对 Colon Tumor 数据集进行特征选择的遗传算法迭代运行结果，图中横坐标代表遗传算法的迭代次数，纵坐标代表每一代种群得到的最优结果(即最低的感知器分类错误率)。从图 2(a)中可以看出，在遗传算法的迭代过程中，2 维 Colon Tumor 数据集的线性分类错误率在持续下降，最后收敛到一个较低的错误率 8.06%。而 3 维的 Colon Tumor 数据集则具有更好的可分性，可收敛到一个更低的分类错误率 3.23%。图 3 给出了 Colon Tumor 数据在 FSIGGA 算法所选出的优化的 2 维(图 3(a))及 3 维(图 3(b))特征子空间中的分布散点图，分别用“*”及“▽”代表两类样本，从图中可看出样本集在对应的特征子空间中具有较好的线性可分性，其中 3 维特征子空间中的线性可分性要优于 2 维特征子空间。图 4 给出了不同特征维数下的分类错误率，从图中可以看出，随着维数的增加，针对 FSIGGA 算法所选出的近似最优特征子空间的感知器分类错误率在下降，且维数达到 7 以后，分类错误率降至 0。

结论 本文主要针对高维数据的特征选择问题，融合 filter 及 wrapper 特征选择模型，提出了一种基于信息增益及遗传算法的特征选择算法。实验证明，该算法能较为有效地找出具有较好的可分离性的特征子集，从而实现降维并提高分类精度。

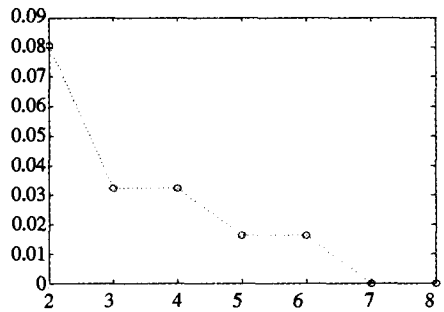

图 4 不同特征维数下的分类错误率

参 考 文 献

1 John G H, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. In: Proc. of the Eleventh Intl. Conf. on Machine Learning, 1994. 121～129

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

3 Liu Huan, Yu Lei. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(5): 491～502

4 Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 1998, 13(2): 44～49

5 YU Lei, Liu Huan. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, 2004(5): 1205～1224

6 Mitra P, Murthy C A. Pal S K. Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301～312