基于遗传算法及聚类的基因表达数据特征选择*)

任江涛　黄焕宇　孙婧昊　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(如基因表达数据)的特征选择,一方面可以提高分类及聚类的精度和效率,另一方面可以找出富含信息的特征子集,如发现与疾病密切相关的重要基因。针对此问题,本文提出了一种新的面向基因表达数据的特征选择方法,在特征子集搜索上采用遗传算法进行随机搜索,在特征子集评价上采用聚类算法及聚类错误率作为学习算法及评价指标。实验结果表明,该算法可有效地找出具有较好可分离性的特征子集,从而实现降维并提高聚类及分类精度。

关键词　特征选择,遗传算法,聚类,基因表达数据

Gene Expression Data Feature Selection Based on GA and Clustering

REN Jiang-Tao　HUANG Huan-Yu　SUN Jing-Hao　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 such as gene expression data, feature selection not only can improve the accuracy and efficiency of classification and clustering, but also can discover informative feature subset, such as genes highly related to some diseases. This paper proposes a new feature selection method for the gene expression data, which realizes the feature subset search by genetic algorithm, and the feature subset is evaluated by the clustering algorithm and the error rate. The experiments show that the proposed algorithm can find the feature subsets with good separability, which results in the good clustering and classification accuracy.

Keywords　Feature selection, GA, Clustering, Gene expression data

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]。

随着基因芯片技术的发展及生物信息学研究的深入,采用模式识别及数据挖掘技术对基因表达数据进行分析,引起了越来越多研究人员的兴趣。基因芯片能够同时分析大量的信息,包括单核苷酸变异多态性、已表达序列标志和基因克隆等。用基因芯片测定细胞生长不同时期的基因表达、测定正常组织与肿瘤组织的 DNA 变化、测定用药前后 DNA 发生的变化、测定基因突变等,就可能发现新药、进行疾病的基因诊断、疾病的预报、弄清人类生物学的奥秘,因此芯片的数据分析显得尤为重要。基因芯片数据分析主要通过对芯片各点数据的比较分析和芯片间的数据比较来实现,其中基因表达数据的分类及聚类是重要的研究方法之一。由于基因表达数据的特征数目很大,成千上万,因此在基因表达数据的聚类及分类问题中特征选择问题尤为突出。该问题的解决可产生两个方面的效益:一方面可以通过降维提高分类及聚类的精度及效率,如在基于基因表达数据的疾病诊断领域可降低误诊率;另一方面找出的特征子集含有丰富的信息,如可帮助人们发现导致疾病的重要基因[3]。

针对上述高维基因表达数据的特征选择问题,本文提出了一种基于遗传算法及 K 均值聚类的特征选择方法。在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。同时,在特征子集评价上采用聚类算法及聚类错误率作为学习算法及评价指标,避免了传统的基于分类器学习算法的复杂过程。

本文第 2 部分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　相关工作

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

*)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。

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[5~7]。

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。由于分类器的训练及测试步骤较多，因此算法流程繁琐。实际上，在聚类对象为有类别标签的数据时，聚类本身也是一种分类方法，也可以实现特征子集的评估，并且其流程简单，易于实现[4,9]。因此，本文采用聚类算法及聚类错误率作为学习算法及相应的评价指标。由于聚类算法是一种机器学习方法，因此所提出的算法是一种 wrapper 型的特征选择算法。

3 算法描述

根据上述讨论，本文提出了一种基于遗传算法及聚类的特征选择方法。与传统的基于遗传算法的特征选择方法相比，本方法主要在编码方案以及基于聚类结果的评价函数定义等方面有所不同，因此下面主要从这两个方面进行阐述，最后给出整个算法的流程。

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；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 Breast Cancer 的维数为 24481，采用直接二进制的编码方法就需要长度为 24481 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 $j<<d$ 时，可得到较小的二进制串。同时，可保证每次选择的特征个数是一致的，从而实现了对特征数量的灵活控制。

3.2 适应度定义

在大多数基于遗传算法的特征选择方法中，都采用不同的分类器模型对所选择的特征集合进行评价。首先基于所选出的特征子集对原特征空间进行降维，形成新的样本集。然后将样本集分为训练集及测试集，先利用训练集训练分类器模型并得到相应的模型及参数，接着输入测试集并对测试结果进行评价，一般利用评价得到的分类精度作为适应度函数。一般来说，训练分类器的代价较高，且上述过程较为复杂。

近年来，面向聚类算法的无监督特征选择方法的研究也取得了很多成果，人们发现特征子集的优劣同样对聚类效果有很大的影响。聚类算法的评价方法之一就是通过对带有类别标签的数据进行聚类，然后统计聚类的错误率，作为聚类算法性能的一种度量。因此，也可以通过采用聚类算法及相应的评价给出特征子集的评价函数。在本研究中，采用 K 均值算法及聚类错误率作为评价方法及适应度。通常情况下，整体的聚类错误率为各类错误率的均值，但该度量往往掩盖了部分精度很差的类的指标。为了避免该问题，采用各类错误率的最大值作为整体聚类错误率。评价算法 *Evaluation* 的流程由算法 1 给出。

算法 1 $Evaluation(D,F_s,k)$

输入：数据集 D，特征子集 F_s，聚类个数 k

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 K 均值算法对数据集 D_F 进行聚类，分为 k 类；

3）根据类别标签信息及聚类结果，分别统计每类的错误率 $err_i(i=1,2,\cdots,k)$；

4）令 $err=\mathrm{Max}(err_i),i=1,2,\cdots,k$

5）输出聚类错误率 err 作为特征子集的评价指标，即适应度。算法结束。

3.3 算法流程

根据 3.1 及 3.2 节的讨论结果，基于标准遗传算法框架，得到一种新的基于遗传算法及聚类的特征选择方法，算法具体描述如下。

算法 2 $GAFL(D,F,fn,k,MaxI)$

输入：数据集 D，特征集合 F，选择的特征数 fn，聚类个数 k，最大迭代次数 Max$I$

输出：优化的特征子集

步骤：

(1)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；

(2)根据编码方案，将个体的二进制表达转化为原特征集合中的特征编号，根据这些特征编号进行特征选择，形成特征子集 F_s；

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

(4)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；

(5)根据适应度执行选择操作；

(6)执行交叉操作；

(7)执行变异操作；

(8)返回步骤(2)。

4 实验研究

为了评估上述 GAFL 算法的有效性，采用了两个基因表达数据集 Breast Cancer 和 ALL-AML Leukemia 进行测试。Breast Cancer 数据集有 78 个样本，其中 34 个样本来自经过治疗后在 5 年内旧病复发的乳腺癌患者，另外 44 个样本来自经过治疗后在 5 年之内一直健康而没有发病的乳腺癌患者。数据集含有 24481 个基因，即特征数为 24481。ALL-AML Leukemia 数据集有 38 个骨髓基因样本，其中 ALL 型样本有 27 个，AML 型样本有 11 个，该数据集的特征维数为 7129。

图 1 给出采用 GAFL 算法对上述两个数据集进行特征选择的实验结果，图中的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（即最低的聚类错误率）。在实验中为 Breast Cancer 数据集及 ALL-AML Leukemia 数据集设定的参数 fn（选择特征数）为 5 和 20。从图中可以看出，在遗传算法的迭代过程中，Breast Cancer 数据集的聚类错误率在持续下降，最后收敛到一个较低的错误率 18.75%。而 ALL-AML Leukemia 数据集具有更好的可分性，迭代到第 20 次时错误率就已降至 0，即聚类精度为 100%，且在往后的迭代过程中始终保持 100%的聚类正确率。

结论 本文主要针对高维基因表达数据的特征选择问题，提出了一种基于遗传算法及聚类的特征选择算法，在编码方式、评价方法及适应度定义等方面对传统的基于遗传算法的特征选择方法进行了改进。实验证明，该算法能较为有效地找出具有较好可分离性的特征子集，从而实现降维并提高聚类及分类精度。

（下转第 224 页）

图3 一般方法

图4 预处理-修正方法

实验中的其他压缩性能可参考表1。

表1 粗糙域块分类时的压缩性能比较

	压缩率	PSNR	时间(分)
一般压缩方法	9.2	24.47	1.33
预处理-修正压缩	11.3	28.15	1.67

这里压缩比定义为：图像压缩比＝源图像字节数/压缩编码字节数。图像(每个像素8字节)的信噪比定义为：

$$PSNRE=10\log_{10}\frac{255^2}{\|u-u^*\|^2}$$

其中 u^* 为源图像，u 为压缩解码图像。

表2 精细域块分类时的压缩性能比较

	压缩率	PSNR	时间(分)
一般压缩方法	9.2	26.8	3.16
预处理-修正压缩	11.6	29.2	3.37

考虑细分的情形时，我们采用16×16大小的分块来补充原有域块，初始分类块和细分时的分类块的大小仍分别为8×8和4×4。有关的试验结果见表2。其中预处理-修正模式下，我们仍然根据误差图像的分布情况对分块的位置进行了校正。

结论 基于IFS的图像压缩算法本质上依赖于图像中的自相似性，对具有典型分形特征的图像可以获得非常高的压缩比。在应用中，不同图像的自相似性程度和方式可以不同，通常需要在压缩过程中针对具体图片进行适当的人工干预，以获得最佳的效果。我们提出的基于预处理-修正模式的分形图像压缩方法为压缩过程中出现粗糙编码时提供了一条有效的修正方案，同时在压缩过程中可根据误差图像的分布特征进行一定的人工干预，为获得理想的压缩效果提供了一种思路。

参考文献

1 Barnsley M F. Fractals Everywhere [M]. San Diag, CA: Academic Press, 1988
2 Jacquin A, Barnsley M F. A fractal theory of iterated Markov operation with application to digital image coding: [Ph. D thesis]. Georgia Institute of Technology, 1989
3 Jacquin A, Barnsley M F. Image coding based on a fractal theory of iterated Markov operators, Part 1: Theoretical foundation [R]: [Technical Report]. 91389-016. Georgia Institute of Technology, 1989
4 Jacquin A, Barnsley M F. Image coding based on a fractal theory of iterated Markov operators, Part II: Construction of fractal codes for digital image [R]: [Technical Report]. 91389-017. Georgia Institute of Technology, 1989
5 Jacquin A. Image coding based on fractal theory of iterated contractive trunsformtion [J]. IEEE Trans Image Proc, 1992, 1: 18～30
6 吴子文，吴鹏晖. 一种基于子带分解的分形图像压缩新方法[J]. 微型计算机应用，1998，19(2)：73～75
7 范策. 分形图像压缩的主池预缩小方法. 计算机工程，2001，27(11)：138～140
8 谭郁松，周兴铭，Fason：一种图像快速分形压缩的改进算法[J]. 计算机工程与科学，2004，26(1)：34～37

(上接第156页)

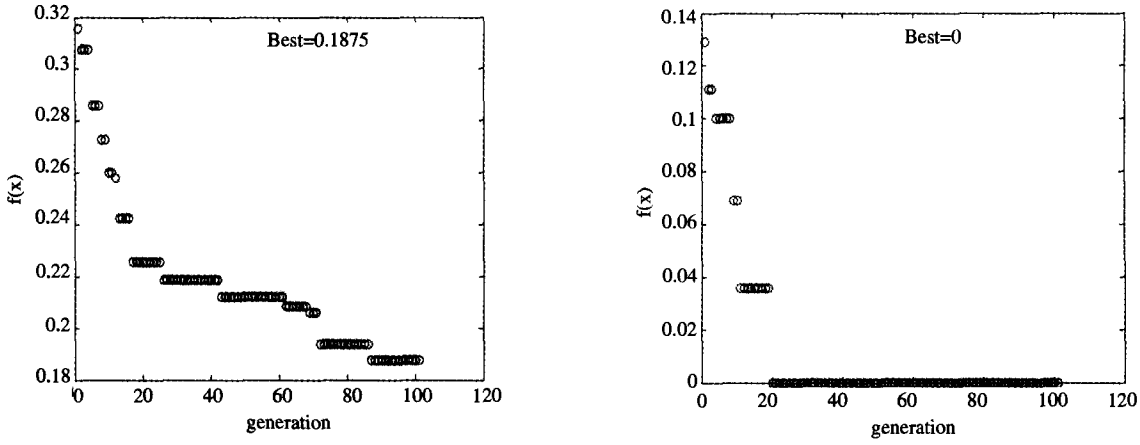

图1 GAFL算法针对Breast Cancer(左图)及ALL-AML Leukemia(右图)数据集的运行结果图

参考文献

1 John G H, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. In: Proceedings of the Eleventh International Conference on Machine Learning, 1994. 121～129
2 Kohavi R, John J H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1-2): 273～324
3 Jiang Daxin, Tang Chun, Zhang Aidong. Cluster analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, 2004, 16(11): 1370～1386
4 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
5 Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 1998, 13(2): 44～49
6 Bhanu B, Lin Yingqiang. Genetic algorithm based feature selection for target detection in SAR images. Image and Vision Computing, 2003, 21(7): 591～608
7 Oh Il-Seok, Lee Jin-Seon, Moon Byung-Ro. Hybrid Genetic Algorithms for Feature Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(11): 1424～1437
8 Han Jiawei, Kamber M. 数据挖掘概念与技术. 范明，孟小峰译. 北京：机械工业出版社，2001
9 Xu Rui, Wunsch II D. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 2005, 16(3): 645～678