基于DNA微阵列数据的癌症分类问题研究进展

于化龙 顾国昌 赵 靖 刘海波 沈 晶
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

摘 要 应用DNA微阵列数据对癌症进行诊断与分型,已经逐渐成为生物信息学领域的研究热点之一。首先概述了基于微阵列数据的癌症分类问题的研究现状与发展趋势。然后简要介绍了微阵列实验的基本步骤,微阵列数据的结构、特点以及用于癌症分类的基本流程。接下来重点从数据预处理、特征基因选择、分类器设计以及分类性能评价等几方面对近10年来的研究成果进行了详细的综述与比较分析。最后,对该领域目前仍然存在的问题进行了归纳并对未来可能的研究方向作出了预测与展望。

关键词 微阵列数据,癌症分类,数据预处理,特征基因选择,分类器设计,分类性能评价
中图法分类号 TP391 **文献标识码** A

State of the Art on Cancer Classification Problems Based on DNA Microarray Data

YU Hua-long GU Guo-chang ZHAO Jing LIU Hai-bo SHEN Jing
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract Applying DNA microarray data to diagnose for cancer and recognize different subtypes of the same tumor has been becoming one of hot topics in Bioinformatics. Firstly, this paper summarized the state of the art and the development trend for cancer classification based on microarray data. Then the basic procedure of DNA microarray experiments, structure & characteristics of microarray data and general process for cancer classification based on DNA microarray data were introduced. After that, a detailed survey and systemic comparative analysis combined with the research results for the last ten years was made from several main aspects listed as below: data preprocessing, feature gene selection, classifier design & classification performance evaluation. Finally, some subsistent difficulties in this research field were summarized and meanwhile, the possible directions for future work were also predicted and suggested.

Keywords Microarray data, Cancer classification, Data preprocessing, Feature gene selection, Classifier design, Classification performance evaluation

DNA微阵列(DNA Microarray),又名基因芯片(Gene Chip),是上世纪末分子生物学领域的一项重大技术突破,它的出现使生物学家希望同时检测成千上万个基因在生物体内活性的梦想成为现实。近年来,DNA微阵列技术已成为后基因组时代进行生命科学研究的基本工具,并被广泛应用于生物学和医学研究的各个领域,如大规模DNA测序[1]、基因多态位点及基因突变检测[2,3]、癌症的诊断与分型[4-8]、基因调控与互作关系挖掘[9,10]以及新药开发[11]等。

在以上的应用领域中,癌症的诊断与分型问题最吸引研究人员的眼球[12]。最早的相关研究可以追溯到1999年,Golub等在著名的Science杂志上发表的题为"Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring"的文章,开创了采用DNA微阵列数据研究癌症分类问题的先河[4]。自此后近10年来,很多生物学、计算机科学、医学甚至统计学领域的研究人员已经被吸引到这一研究领域并根据各自的领域知识提出了大量实用有效的方法与技术,使得该研究方向逐渐发展成为了生物信息学领域的研究热点之一。

国外在这一领域的研究开展得比较早,其中麻省理工学院[4]、普林斯顿大学[5]、斯坦福大学[6]以及哈佛大学[7,8]等多所国外著名学府都在这一领域取得了丰硕的研究成果,它们发表的文献已经被其它研究人员广为引用。相比之下,国内进入这一研究领域相对较晚,直到2005年才逐渐有一些相关的文献发表出来,其中清华大学[13,14]、中国科技大学[15,16]、国防科技大学[17]、北京工业大学[18,19]、哈尔滨工业大学[20]以及哈尔滨医科大学[21]等一批国内研究机构都在该领域进行了探索并开发出了很多实用的方法与技术。为了反映该领域的发展趋势,我们采用{microarray data ∪ gene expression} ∩

到稿日期:2009-11-18 返修日期:2010-01-29 本文受国家自然科学基金(60873036),国家教育部博士点基金新教师项目(20070217051),中国博士后基金(20060400809),黑龙江省青年科技专项资助项目(QC06C022)资助。

于化龙(1982—),男,博士生,CCF学生会员,主要研究方向为生物信息学、模式识别等,E-mail:yuhualong@hrbeu.edu.cn;顾国昌(1946—),男,教授,博士生导师,CCF高级会员,主要研究方向为图像处理、模式识别等;赵 靖(1972—),女,博士,副教授,主要研究方向为机器学习、可信计算与移动计算;刘海波(1976—),男,博士,副教授,主要研究方向为模式识别与图像处理等;沈 晶(1969—),女,博士,副教授,主要研究方向为机器学习、图像处理等。

{cancer ∪ tumor} ∩ classification 作为主题词在 ISI Web of Knowledge 中进行了检索，检索的近 10 年文献发表情况如图 1 所示。

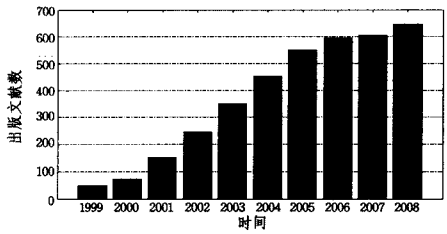

图 1　近 10 年文献发表情况

从图 1 中可以看出，从 1999 年至 2008 年，该领域发表论文的数量呈现出不断快速增长的趋势，表明该领域正在吸引着越来越多科研人员的兴趣。其中 Golub 的那篇开山之作[4]被引用了 3948 次，且随着时间的推移，这一引用数字仍然在不断地攀升。

在以上研究背景下，本文详细综述了基于微阵列数据的癌症分类问题的研究现状以及目前所面临的主要问题，以便为该领域及相关领域的研究者提供有益的参考。本文第 1 节首先介绍了与该领域相关的背景知识，包括微阵列实验的步骤，微阵列数据的结构、特点以及用于癌症分类的基本流程。第 2 节从数据预处理、特征基因选择、分类器设计以及分类性能评价等 4 方面对近 10 年来的研究成果进行了详细的综述与分析。第 3 节总结了该领域目前仍然存在的问题，并对未来可能的研究方向作出了展望。

1　背景知识

DNA 微阵列之所以被称为基因芯片，主要是因为该项技术的概念来源于计算机芯片，它借用了计算机芯片高度集成化的特点，将核酸密集有序地排列在固相载体预先设定的区域内，形成微型的检测器件[12]，检测所获得的数据反映的是基因转录产物 mRNA 在细胞中的丰度。目前根据制备方法的不同，主要可以将基因芯片分为两类：cDNA 微阵列（cDNA microarrays）与寡核苷酸微阵列（Oligonucleotide microarrays）。尽管制备方法不同，但二者的原理是相同的，即利用 4 种核苷酸（A，T，C，G）两两互补配对的特性，使两条在序列上互补的单核苷酸链形成双链，即杂交（Hybridization）过程。

DNA 微阵列实验的基本过程如下：首先要准备好参考样本与测试样本的组织细胞，并从这些细胞中分别提取 mRNA 分子；然后使用反转录酶将 mRNA 转录为 cDNA，并使用不同的荧光素对参考样本（绿色荧光素 Cy3）与测试样本（红色荧光素 Cy5）分别进行标记；接下来对两标记后的 cDNA 样本进行纯化并混合，将混合样本与微阵列探针进行充分的杂交；最后，采用激光或荧光显微镜检测杂交后的芯片，并使用图像处理与分析软件来获取基因表达的强度信息。通常，采用两种荧光强度的对数比来表示基因表达值[22]，如式（1）所示：

$$\text{gene\_expression}=\log_2\frac{\text{Int(Cy5)}}{\text{Int(Cy3)}} \tag{1}$$

在以上描述的实验过程中，将测量的每个对象称为一个基因，每次实验称为一个样本，则一个 DNA 微阵列数据集（也可以称为基因表达谱）就是由在 n 次实验下所测得的 m 个基因的表达值所组成，可以用 $m\times n$ 维矩阵来表示。矩阵中第 $i(1\leqslant i\leqslant m)$ 行、第 $j(1\leqslant j\leqslant n)$ 列元素所对应的数值代表在第 j 次实验中所测得的第 i 个基因的表达值，可以用 x_{ij} 来表示。

由于微阵列实验的独特性，导致微阵列数据集具有如下一些特点：

①小样本：微阵列实验的成本通常是很昂贵的，如 Affymetrix 公司生产的 2.5 万全基因组芯片，价格大约为 1000 美元，即使定制一个约 200 条基因的 pathway 芯片也需要支付至少 200 美元[12]。高昂的测试费用限制了实验的次数，导致了微阵列数据集小样本的特性。一般常见的微阵列数据集都只包含几十到几百个样本不等。

②高维数：微阵列实验的独特之处就在于它可以同时测试成千上万个基因的表达情况，它的这一特性也导致了微阵列数据集高维数的特点。通常，一个微阵列数据集包含几千到两万个基因不等。

③高噪声：由于在微阵列实验中会不可避免地出现系统误差甚至人为失误，从而导致了微阵列数据集中含有大量的噪声数据，这些数据既包括一些不准确的极端数据，也包括那些由于某种原因而缺失了的数据。

④样本分布不均衡：在很多微阵列数据集中都会出现某一类或某几类的样本数明显多于其它类别的情况，这种样本分布不均衡的现象将会对后期的数据分析带来很多负面的影响[20]。

由此可见，微阵列数据集具有高维小样本、高噪声以及样本分布不均衡等特点，它的这些特点注定了它与其它数据载体的不同。与传统的分类问题相比，采用微阵列数据对癌症进行分类更加困难，因为它具有一套独有的流程，如图 2 所示。

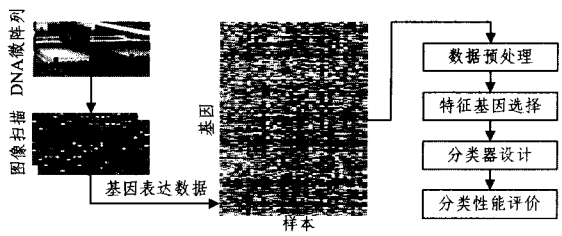

图 2　微阵列数据获取与分类的一般过程

2　研究现状

本节将从数据预处理、特征基因选择、分类器设计以及分类性能评价等 4 个主要方面对近 10 年来的研究成果进行综述与归纳分析。

2.1　数据预处理

2.1.1　缺失值填补

很多微阵列数据集都含有或多或少的缺失值。导致这些缺失值的原因可能有很多，如芯片擦伤、灰尘、杂交失败以及图像污染等。这些缺失值的存在将会对后期的数据分析产生严重的负面影响，如聚类的效果以及差异表达基因选取的准确性等[23,24]。可以通过重做实验来解决以上问题，但是由于成本较高，使其基本不可行。因此只能通过利用数据集内部的结构关系对缺失值进行填补来尽可能恢复数据的本来面目。

最早在文献中提出应对微阵列缺失数据进行填补的是

Alizadeh[6]。他提出的方法也很简单，即使用0值或行均值来填补缺失值。这两种方法尽管有较低的时间复杂度，但是并没有考虑数据的结构关系，因而导致填补后的数据集与真实数据集之间存在较大的偏差。为此，Troyanskaya 等[25]提出了两种新的方法，分别为 K 近邻填补法（KNNimpute）和奇异值分解填补法（SVDimpute）。实验结果显示，KNNimpute 对缺失数据的比例不敏感，并且在非时间序列和含有大量噪声的数据集上表现得更好，而 SVDimpute 则在时间序列数据集上有比较好的效果。Oba 等[26]提出了贝叶斯主成分填补法（BPCAimpute），该方法在大多数数据集上的实验结果都要好于前面提到的方法，且有着对参数、噪声以及数据缺失比例不敏感的优点，但不适合有明显局部相似结构的数据集。针对它的这一缺点，Kim 等[27]提出了局部最小二乘填补法（LLSimpute），实验证明该方法不但善于发现数据中的局部相似结构，而且有很好的鲁棒性，是一种有竞争力的方法。文献[28]提出了支持向量回归填补法（SVRimpute），该方法由于使用了一种正交输入编码方案，因而充分利用了基因表达矩阵中缺失值的信息，其性能通常好于其它方法，但是在数据集中缺失值比例较高时，它的性能要低于 BPCAimpute 以及 LLSimpute；其另一个缺点是时间复杂度高。Tuikkala 等[29]利用 Gene Ontology 数据库中提供的基因语义相似性注释信息来重构缺失的基因表达值（GOimpute），但该方法只在缺失值比例较高时才能体现出优势。

文献[30,31]对现有的经典缺失值填补方法进行了大量的实验比较分析，发现它们都有各自的优点与适用范围，没有哪一种方法具有绝对的优势。

2.1.2 数据标准化

数据预处理的另一个任务是对数据进行标准化。这主要是因为后期的数据分析常常要求各个基因要处于同一量纲之下，尤其是那些基于欧氏距离的数据处理方法更是如此。现有常用的微阵列数据标准化方法主要有以下两种：

（1）限制每个基因的均值为0、方差为1，计算方法如下：

$$x_{ij}^{*}=(x_{ij}-\overline{x_i})/s_i \tag{2}$$

式中，x_{ij} 与 x_{ij}^{*} 分别为第 j 个样本在第 i 个基因上的原始与变换后的表达值，而 $\overline{x_i}$ 与 s_i 则分别表示第 i 个基因在所有样本上的平均值与标准差。

（2）限制每个基因的值在[0,1]范围内，计算方法如下：

$$x_{ij}^{*}=(x_{ij}-x_{\min})/(x_{\max}-x_{\min}) \tag{3}$$

式中，$x_{\max}$ 与 $x_{\min}$ 分别表示第 i 个基因在所有样本上的最大与最小值。

几乎现有的全部文献都是采用以上两种方法之一来对微阵列数据进行标准化的。

2.2 特征基因选择

由于微阵列数据集具有高维小样本、高噪声、高冗余等特点，从而导致了“维数灾难”和“过适应”现象的出现[32]。事实上，数据集中只有很少一部分基因与分类任务是密切相关的，这些基因被称为特征基因，而选取这些特征基因的过程便被称为特征基因选择。对特征基因进行选择主要有以下3方面的意义：

①降低临床诊断的费用；

②提高临床诊断的精度；

③为从分子层上研究致病机理和发现新的药物靶点提供便利。

近年来，特征基因选择已经成为癌症分类研究领域中最为热门的方向之一。研究者们从不同角度提出了大量的特征基因选择方法。国内的研究者们也对这些研究成果进行了比较全面与详细的综述[33,34]，但他们对方法的归纳与分类稍显凌乱，因此有必要对这一问题进行更清晰与深入的讨论。传统的观点通常将特征基因选择方法分为两大类：过滤法（Filter）与缠绕法（Wrapper）[35]，二者的最大区别在于在选择特征的过程中是否存在分类器的参与。二者相比，通常是过滤法的速度更快，而缠绕法的性能更好。为了包容以上二者的优点，近年来提出了越来越多的混合方法，这些方法既包括过滤法与过滤法的混合[18,36]，也包括过滤法与缠绕法的混合[37,38]，这类方法已经成为了当今研究的趋势。

2.2.1 过滤法

a）基于排序（Ranking）的方法

基于排序的方法是最为简单与常用的一类特征基因选择方法，它通常采用某种策略来评估各个基因对分类任务的重要性，然后按照重要性高低对基因进行排序，最后选择一定数量排名靠前的基因作为特征基因。在这类方法中，“信噪比”评价指标[4]是最为常用的，它评价的是每个基因在所有样本上的类间松散度与类内紧密度之比。比例越高，该基因也就越重要。相似的方法还有“t-检验”[39]、“Fisher 指标”[40]以及“分类信息指数”[18]等。Dudoit 等[41]将这种方法的思想进行了扩展，并将其应用于多类问题中。通常，对分类任务贡献度最大的应该是那些与类别标签关系最为密切的特征。Cho 等[42]从这一角度出发分别采用 Pearson 相关系数、余弦相似度以及欧氏距离来测试各个基因与类别标签的相似程度，并选择相似度高的基因作为特征基因。另外，信息增益（Information Gain，IG）[38]，Relief[19]，马尔可夫毯（Markov Blanket）[43]以及质心收缩（shrunken centroids）[44,45]等方法也在特征基因选择领域得到了应用。

基于排序的方法具有较小的计算开销，但它将基因简单地割裂开，而没有考虑它们之间的相互作用。事实上，被排序方法排除的一些基因也含有很多的分类信息，不应该简单地将它们排除掉[46]。

b）基于聚类（Clustering）的方法

这类方法通常考虑到微阵列数据集中基因高度冗余的特点，首先通过聚类方法将相互冗余的基因聚集到同一个簇，然后从每个簇中选取一个代表来组成最终的特征基因集合。文献[47-49]都是采用这种思想来进行特征基因选择的。在选取代表基因时，既可以使用排序方法选出各簇中与分类关系最为密切的基因[47,48]，也可以使用簇内基因的均值[49]。很明显，基于聚类的方法由于考虑到了冗余基因的影响，因而通常可以获得比排序法更好的性能。

2.2.2 缠绕法

a）基于启发式搜索（Heuristic Search）的方法

启发式搜索是最为常见的缠绕型特征选择方法，它通常采用分类器直接对一个特征基因子集进行分类性能评价，然后根据评价的结果采取某种策略对子集进行调整，以达到不断探索最优子集的目的。遗传算法（Genetic Algorithm，GA）则是最为常用的启发式搜索算法之一。Li 等[50]结合 GA 与 KNN 分类器选取特征基因子集；Hong 等[51]修改了 GA 中染

色体的表达方式，从而解决了大规模特征基因选择问题中存在的困难；文献[52]将 GA 与 SVM 分类器相结合，并创新性地采用支持向量的距离作为适应度函数，取得了很好的效果。除了 GA 以外，很多其它常用的启发式搜索方法也在这一领域得到了应用，如顺序前进法(Sequential Forward Selection，SFS)[35]、粒子群算法(Particle Swarm Optimization，PSO)[53]以及蚁群算法(Ant Colony Optimization，ACO)[54]等。

基于启发式搜索的特征基因选择方法具有选取特征少、分类性能好等优点，但也有一个不可回避的缺点，即时间复杂度非常高。

b)基于递归约减(Recursive Elimination)的方法

这类方法是由 Guyon 等[55]于 2002 年提出的，她采用支持向量机作为分类器，首先在整个数据集上测试分类的性能，然后计算移出每个基因后的性能变化，选择分类函数中关联权重绝对值最小的基因，并将其从训练集中移去，重复此过程直至训练集数据为空，最后一步删除的特征基因子集就是最优分类子集。Zhou 等[56]在以上方法的基础上做出了改进，提出了 MSVM-RFE 方法，并成功地将其应用于多类问题。总的来说，基于递归约减的方法有很好的分类性能，但是时间复杂度非常高。

c)基于集成(Ensemble)的方法

为了获得与癌症最为相关的特征基因，Li 等[21,57]设计了一种集成决策特征基因选择方法。该方法首先采用重取样技术生成大量不同的训练与测试集，然后以分类错误率最小为评价标准，重复运行递归分层特征基因选择方法，最后综合集成各特征选择器的结果，以生成最终的特征基因子集。这类方法的特点是选出的特征基因子集未必对分类任务最优，但一定与癌症存在着最为密切的关系，其缺点在于有较高的时间复杂度。

2.2.3 *混合法*

a)过滤法＋过滤法

这种方法通常是用第一个过滤法来过滤分类无关基因，用第二个过滤法来去除冗余基因。如 Wang 等[36]提出的 HykGene 方法，就是首先使用某种基因排序方法来进行初选，再使用层次聚类的方法来去除冗余。文献[18]中提出的方法与 HykGene 的不同之处在于使用了 Pearson 相关系数来消除冗余基因。这类方法的优点是分类精度高，选取的特征基因数少，且时间复杂度也不高。

b)过滤法＋缠绕法

这类方法首先使用过滤法进行特征基因的初选，然后在这些选出的基因上使用缠绕法探索最优的特征基因集。文献[37]首先利用 3 种性质不同的排序法来构造候选特征基因池，然后在其上采用 GA 进行搜索。实验结果显示该方法可以通过选出的很少的基因，来获得很高的分类准确率。文献[38]的实验结果也表明了该类方法的有效性。

各类特征基因选择方法都有各自的优缺点，没有哪种方法具有绝对的优势。在实际应用中，应根据具体情况来选择适合的方法。

2.2.4 *特征提取方法*

另外，一些经典的特征提取方法也在微阵列数据降维方面得到了应用，包括主成分分析(Principle Component Analysis，PCA)[58]、小波变换(Wavelet Transform，WT)[59]等。这类方法的优点在于在降低数据集维数的同时，最大限度地保留了原始数据集中的信息。但提取的特征由于是变换后的结果，无法从生物学与医学角度进行解释，因此并不实用。

2.3 分类器设计

分类器的选取与设计是癌症分类问题的核心，也是决定分类任务最终成败的关键。根据分类策略与任务的不同，主要可以将其分为以下 3 类。

2.3.1 *单一分类方法*

近年来，很多经典的分类方法都在微阵列数据分类领域得到了应用，如决策树(Decision Tree，DT)[60]、K 近邻分类器(K-nearest neighbors，KNN)[50]、人工神经网络(Artificial Neural Networks，ANN)[61]、线性贝叶斯分类器(Linear Bayesian Classifier，LBC)[62]以及支持向量机(Support Vector Machine，SVM)[17,40]等。文献[63]对这些分类方法在大量数据集上进行了测试，得出了 SVM 的性能要普遍优于其它分类方法的结论。

2.3.2 *集成分类方法*

最近几年，研究人员开始更多地关注分类精度更高、泛化能力更强的集成分类方法。众所周知，构造一个成功的集成分类器需要从两方面着手：一是要保证各个基分类器的准确率，二是要尽量增加它们之间的差异。前者可以通过选用较优的分类方法予以保证，因此如何获取基分类器间的差异便成为了研究人员关注的焦点。现有的文献也主要是围绕着这一问题进行展开的。

根据获取差异时所使用策略的不同，可以将现有的集成分类方法分为以下两类：

(1)构造性集成分类方法

这类方法的思想较为简单，即构造大量的基分类器，并采用某种策略来保证这些基分类器间的差异。一些比较著名的集成分类方法，如 Bagging、Boosting 以及随机森林(Random Forest)等都属于此类。2003 年，Tan 等[64]第一次将集成分类方法应用于微阵列数据分类领域，他们测试了单决策树、Bagging 决策树以及 AdaBoost 决策树的性能，得出了集成方法可以提高微阵列数据分类精度的结论；Dettling 等[65]结合 Bagging 与 Boosting 的优点提出了 BagBoosting 方法，并在大量肿瘤微阵列数据集上对其有效性进行了验证；文献[66]针对微阵列数据特征高度冗余这一特性，将随机子空间(Random Subspace)集成分类方法成功应用于癌症分类问题；Diza-Uriarte 等[67]采用 9 个微阵列数据集测试了随机森林分类器的性能，得出了其分类精度高于 SVM 的结论，但这一结论受到了 Statnikov 等[68]的质疑；Liu 等[15]将旋转森林(Rotation Forest)分类方法与多个不同的特征提取技术相结合，并进行了大量的实验，结果显示该方法只需集成很少的基分类器，便可以获得很高的分类准确率。

(2)选择性集成分类方法

根据 Zhou 等[69]的思想可知，如果采用某种策略从大量的基分类器中选出一些差异度较大的个体进行集成，预测的结果可能更加准确。基于选择性差异的集成分类方法便是受到了这种思想的启发。Peng[70]首先采用 K-means 方法将判别空间相似的基分类器进行聚类，然后在每一类中选择一个代表进行集成，有效地增加了基分类器间的差异性，从而达到了提高分类精度的目的；文献[61,71]则采用启发式搜索的方

法来寻找最优的分类器组合。

以上两类方法相比，前者的分类性能稍差，集成规模也较高，但存在着较低计算代价的优势，而后者的时间复杂度通常较高。另外，值得注意的一点是，后者通常是建立在前者之上的。

2.3.3 多类分类方法

传统的微阵列数据分类方法主要是为二类问题而设计的，无法将其直接应用于多类问题。为此，研究人员提出了一些变通的方法，如 Yeang 等[72]分别采用"一对多"和"一对一"的编码方法将多类问题转化为多个二类问题，并结合 K 近邻、加权投票与支持向量机分类器对性能进行了比较，比较结果显示"一对多"支持向量机(One-Versus-Rest Support Vector Machine，OVR-SVM)可以获得最为突出的分类性能；Lee 等[73]提出了一种多类别支持向量机，可以使用一个分类器直接对多类样本做出判别，其缺点在于参数过多、构造复杂、计算量大；Berrar 等[74]采用基于概念学习的 K 近邻分类器来解决多类分类问题，结果显示这种方法比传统 K 近邻分类器性能更好；Shen 等[75]结合了多种不同的输出码编码与解码准则来构造多类支持向量机，获得了较高的分类精确度；文献[76]则对各种多类别分类方法进行了细致的比较，并得出了 OVR-SVM 分类器分类性能最优的结论。尽管研究者们已经提出了很多实用的方法，但与二类问题相比，多类问题的分类精度还是显得过低，因此在未来它仍将会是这一领域的研究重点。

2.4 分类性能评价

在设计好分类器后，采用何种方法可以对其性能做出真实精确的评价，便成为了摆在研究者面前的一个新难题。在模式识别领域，研究人员通常采用保留法(holdout)来评价分类器的性能，即将原始数据集划分为训练集与测试集，采用训练集进行训练，在测试集上作出评价。其优点在于评价偏差小，同时有很小的计算代价，但仅适用于大样本数据集，并不适合以"小样本"为特点的微阵列数据集。在微阵列数据分类领域，交叉验证(Cross-Validation，CV)的方法更加受到青睐。该方法的思想是将样本集随机等分为几部分，每次从中抽取一部分作为测试样本，而用其它样本对分类器进行训练，当各部分样本都做过一次测试集后，统计错分样本次数与原始样本规模的比，并将其作为分类误差。当将样本集划分为 n 部分时，称之为 n 折交叉验证法；而当 n 等于原始数据集的样本数时，则称为留一交叉验证法(Leave One Out Cross-Validation，LOOCV)。交叉验证法有效地利用了微阵列数据集中的样本，从一定程度上解决了"小样本"数据集评估困难的难题。

为了验证交叉验证法在微阵列数据集上的有效性，Braga-Neto 等[77]对其与其它几种性能评价方法进行了大量的测试与比较，结果发现该方法尽管对错误估计的偏差很小，但仍存在着方差大与离群点多等缺点，其中留一交叉验证法尤为严重。该文献同时也发现置换验证法(resubstitution，即同时使用原始样本集对分类器进行训练与测试)会严重高估分类器的性能，而重取样法(0.632 bootstrap)尽管评估偏差不大，方差也较小，但却需消耗大量的计算资源。文献[78]将重取样法与交叉验证法相结合，提出了一种混合的方法(Bootstrap Cross Validation，BCV)，该方法尽管改善了交叉验证法方差较大的缺陷，但同时大幅度增加了计算的负担。Varma 等[79]总结了大量与微阵列数据分类相关的文献，发现很多研究者都将特征选择的过程或参数优化的过程嵌入到了交叉验证的内部，这两种作法会使分类的错误率大大地被低估，从而导致评价的偏差。为此，Wood 等[80]提出了一种二层外部交叉验证方法(Two-level external Cross-Validation，2-ext CV)，即每次在外部保留一折样本，而在其它样本上选择特征基因与最优参数，这种方法可以有效地消除选择偏差(Selection bias)和优化偏差(Optimization bias)，从而为分类器性能提供更加准确的评价。

表 1 在微阵列数据应用背景下对一些主要的分类性能评价方法的特点进行了总结。

表 1 一些主要的分类性能评价方法比较

Method	Bias	Variance	Computational Cost
resubstitution	High	Low	Low
n-fold CV	Low	High	Medium
LOOCV	Low	High	Medium
0.632 bootstrap	Medium	Low	High
BCV[78]	Low	Low	High
2-ext CV[80]	Low	High	High

正如 Dougherty 在文献[81]结论中所写的那样："Sound science requires conclusions to be drawn only when conclusions are warranted"，研究者只有尽量选择那些低偏的分类性能评价方法，才能保证所发布成果的可靠性与价值。

3 存在的问题与未来研究展望

尽管经过了 10 年的研究与探索，提出了大量相关的方法与技术，但该领域仍远远没有达到成熟的地步，还存在着许多有待解决的问题，主要包括：

(1) 对一些由数据集结构本身引发的问题缺乏关注，如数据集样本分布不均衡[20]以及选取的最优特征基因数与样本数之间的关系[82]等。这些问题都是在现实中普遍存在的且会对分类性能产生极大的影响。因此，研究者有必要对这些问题投入更多的关注。

(2) 缺乏统一的评价标准。研究人员在公布自己方法的时候，都会宣称自己方法是最优的，但在与前人工作做对比时，所采用的测试方法、评价标准往往是不同的，这种不同使他们工作的对比缺乏意义。如何在统一的评价标准下，大范围地测试并比较已有方法在多方面的性能，也是一个迫切需要解决的问题。

(3) 该领域的研究目前仍停留在实验室阶段，缺乏临床医学的佐证。导致这一问题的原因是多方面的，除了现有方法的准确性仍难以让人满意外，微阵列检测的成本过高也是主要原因之一。因此，为了解决这一问题，仍需要多方面继续不懈地努力。

针对微阵列数据分类领域的研究现状与发展趋势，我们认为以下几方面是未来研究中值得关注的。

(1) 微阵列数据集的整合。目前，在互联网上已公布了大量的微阵列数据集，其中很多数据集都是针对同一种癌症做出的检测。如果能将它们有机地融合在一起，便可以有效地解决"小样本"问题，并可以对癌症分类问题给出更加令人信服的结果。研究人员在这一方面已经进行了一些初步的探索[83-84]，但是尚不成熟。

(2) 未标记样本的利用。随着微阵列实验成本的不断下降,研究人员已经可以以同样的开支获得远多于以往的样本,但是对这些样本作出诊断却需要耗费大量的人力与财力。因此,将主动学习[85]引入这一研究领域是完全必要的,它将成为解决这一问题的利器。

(3) 先验生物/医学知识的结合。这一领域的方法大多是由统计学家与计算机科学专家提出的,他们生物/医学知识的匮乏使得这些方法缺少与生物/医学先验知识的结合。相信结合了先验知识的方法将比现有方法更加具备竞争性。

本文针对近年来基于微阵列数据的癌症分类领域的研究成果进行了详细的综述与分析,评价了现有方法的优缺点,归纳了仍存在的亟待解决的问题并对未来的研究方向作出了展望。希望本文可以对该领域及相关领域的研究人员提供有益的参考。

参 考 文 献

[1] Southern E M. DNA chips: analyzing sequence by hybridization to oligonucleotides on a large scale[J]. Trends in Genetics, 1996,12(3):110-115

[2] Hacia J H, Brody L C, Chee M S, et al. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis[J]. Nature genetics, 1996,14(4):441-447

[3] Wang D G, et al. Large-scale Identification, Mapping, and Genotyping of Single-nucleotide Polymorphisms in the Human Genome[J]. Science, 1998,280(5366):1077-1082

[4] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999,286(5439):531-537

[5] Alon U, Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide array[J]. Proceedings of the National Academy of Sciences, 1999,96(12):6745-6750

[6] Alizadeh A A, Eisen M B, Davis R E, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J]. Nature, 2000,403(6769):503-511

[7] Pomeroy S L, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression[J]. Nature, 2002,415(6870):436-442

[8] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nature Medicine, 2002,8(1):68-74

[9] Soinov L A, Krestyaninova M A, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning[J]. Genome Biology, 2003,4(1):R6

[10] Pe'er D, Regev A, Elidan G, et al. Inferring subnetworks from perturbed expression profiles[J]. Bioinformatics, 2001,17(Suppl 1):S215-S224

[11] Evans W E, Guy R K. Gene expression as a drug discovery tool [J]. Nature Genetics, 2004,36(3):214-215

[12] 李瑶. 基因芯片数据分析与处理[M]. 北京:化学工业出版社, 2006

[13] Zhang C L, Lu X S, Zhang X G. Significance of Gene Ranking for Classification of Microarray Samples[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006,3(3):312-320

[14] Zhang X G, Lu X, Shi Q, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data[J]. BMC Bioinformatics, 2006,7:197

[15] Liu K H, Huang D S. Cancer classification using rotation forest [J]. Computers in Biology and Medicine, 2008,38(5):601-610

[16] Huang D S, Zheng C H. Independent component analysis based penalized discriminant method for tumor classification using gene expression data[J]. Bioinformatics, 2006, 22(15): 1855-1862

[17] 王树林,王戟,陈火旺,等. 肿瘤信息基因启发式宽度优先搜索算法研究[J]. 计算机学报,2008,31(4):636-649

[18] 李颖新,阮晓钢. 基于支持向量机的肿瘤分类特征基因选取[J]. 计算机研究与发展,2005,42(10):1796-1801

[19] 李颖新,李建更,阮晓钢. 肿瘤基因表达谱分类特征基因选取问题及分析方法研究[J]. 计算机学报,2006,29(2):324-330

[20] 李建中,杨昆,高宏,等. 考虑样本不平衡的模型无关的基因选择方法[J]. 软件学报,2006,17(7):1485-1493

[21] Li X, Rao S Q, Wang Y D, et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling[J]. Nucleic Acids Research, 2006,32(9):2685-2694

[22] Derisi J L, Iyer V R, Brosn P O. Exploring the metabolic and genetic control of gene expression on a genomic scale[J]. Science, 1997,278(5338):680-686

[23] Jornsten R, Wang H Y, Welsh W J, et al. DNA microarray data imputation and significance analysis of differential expression [J]. Bioinformatics, 2005,21(22):4155-4161

[24] Brevern A G, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering[J]. BMC Bioinformatics, 2004,5:114

[25] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6):520-525

[26] Oba S, Sato M, Takemasa I, et al. A Bayesian missing value estimation method for gene expression profile data[J]. Bioinformatics, 2003,19(16):2088-2096

[27] Kim H, Golub G H, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation [J]. Bioinformatics, 2005,21(2):187-198

[28] Wang X, Li A, Jiang Z H, et al. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme[J]. BMC Bioinformatics, 2006,7:32

[29] Tuikkala J, Elo L L, Nevalainen O S, et al. Improving missing value estimation in microarray data with gene ontology[J]. Bioinformatics, 2006,22(5):566-572

[30] Tuikkala J, Elo L L, Nevalainen O S, et al. Missing value imputation improves clustering and interpretation of gene expression microarray data[J]. BMC Bioinformatics, 2008, 9: 202

[31] Brock G N, Shaffer J R, Blakesley R E, et al. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes[J]. BMC Bioinformatics, 2008,9:12

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

[33] 周昉,何洁月.生物信息学中基因芯片的特征选择技术综述[J].计算机科学,2007,34(12):143-150

[34] 张丽娟,李舟军.微阵列数据癌症分类问题中的基因选择[J].计算机研究与发展,2009,46(5):794-802

[35] Inza I, Larranaga P, Blanc R, et al. Filter versus wrapper gene selection approaches in DNA microarray domains[J]. Artificial Intelligence in Medicine, 2004, 31(2):91-103

[36] Wang Y H, Makedon F S, Ford J C, et al. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data[J]. Bioinformatics, 2005, 21(8):1530-1537

[37] Tan F, Fu X Z, Wang H, et al. A Hybrid Feature Selection Approach for Microarray Gene Expression Data[C]// Proceedings of Sixth International Conference on Computational Science. LNCS3992. Berlin Heidelberg: Springer, 2006:678-685

[38] Chuang L Y, Ke C H, Chang H W, et al. A Two-stage Feature Selection Method for Gene Expression Data[J]. OMICS, 2009, 13(2):127-137

[39] Baldi P, Long A D. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes[J]. Bioinformatics, 2001, 17(16): 509-519

[40] Furey T S, Cristianini N, Duffy N. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906-914

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

[42] Cho S B, Won H H. Machine learning in DNA microarray analysis for cancer classification[C]// Proceedings of First Asia-Pacific bioinformatics conference. Adelaide, Australia, CRPIT, 2003, 19:189-198

[43] Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data[C]// Proceedings of 18th International Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001:601-608

[44] Tibshiranit R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[J]. Proceedings of the National Academy of Sciences, 2002, 99(10): 6567-6572

[45] Shen Q, Shi W M, Kong W. New gene selection method for multiclass tumor classification by class centroid[J]. Journal of Biomedical Informatics, 2009, 42(1):59-65

[46] Czekaj T, Wu W, Walczak B. Classification of genomic data: Some aspects of feature selection[J]. Talanta, 2008, 76(3):564-574

[47] Jaeger J, Sengupta R, Ruzzo W L. Improved gene selection for classification of microarrays[C]// Proceedings of Pacific Symp on Biocomputing. Singapore: World Scientific Publishing Company, 2003:53-64

[48] Au W H, Chan K C C, Wong A K C, et al. Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005, 2(2):83-101

[49] Hanczar B, Courtine M, Benis A, et al. Improving Classification of Microarray Data Using Prototype-based Feature Selection [J]. SIGKDD Explorations, 2003, 5(2):23-30

[50] 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[J]. Bioinformatics, 2001, 17(12):1131-1142

[51] Hong J H, Cho S B. Efficient huge-scale feature selection with speciated genetic algorithm[J]. Pattern Recognition Letters, 2006, 27(2):143-150

[52] Chen X W. Margin-based wrapper methods for gene identification using microarray[J]. Neurocomputing, 2006, 69(16-18): 2236-2243

[53] Chuang L Y, Chang H W, Tu C J, et al. Improved binary PSO for feature selection using gene expression data[J]. Computational Biology and Chemistry, 2008, 32(1):29-38

[54] Robbins K R, Zhang W, Bertrand J K. The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification[J]. Mathematical Medicine and Biology, 2007, 24(4):413-426

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

[56] Zhou X, Tuck D P. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data[J]. Bioinformatics, 2007, 23(9):1106-1114

[57] Yang Y Y, Wang H Y, Li X, et al. A feature ensemble technology to identify moceculer mechanisms for distinction between multiple subtypes of lymphoma[J]. Progress in Natural Science, 2008, 18(12):1491-1500

[58] Li G Z, Bu H L, Yang M Q, et al. Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis [J]. BMC Genomics, 2008, 9(Suppl 2):S24

[59] Liu Y H. Prominent feature selection of microarray data [J]. Progress in Natural Science, 2009, 19(10):1365-1371

[60] Horng J T, Wu L C, Liu B J, et al. An expert system to classify microarray gene expression data using gene selection by decision tree[J]. Expert System with Applications, 2009, 36(5): 9072-9081

[61] Chen Y H, Zhao Y O. A novel ensemble of classifiers for microarray data classification[J]. Applied Soft Computing, 2008, 8(4):1664-1669

[62] Asyali M H. Gene expression profile class predication using linear Bayesian classifiers[J]. Computers in Biology and Medicine, 2007, 37(12):1690-1699

[63] Pirooznia M, Yang J Y, Yang M Q, et al. A comparative study of different machine learning methods on microarray gene expression data[J]. BMC Genomics, 2008, 9(Suppl 1):S13

[64] Tan A C, Gilbert D. Ensemble machine learning on gene expression data for cancer classification[J]. Applied Bioinformatics, 2003, 2(Suppl 3):S75-S83

[65] Dettling M. Bagboosting for tumor classification with gene expression data[J]. Bioinformatics, 2004, 20(18):3583-3593

(下转第 32 页)

A new explanation for the effectiveness of voting methods[C]// Proceedings of the 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1997:322-330

[40] Freimd Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121(2):256-285

[41] Quinlan J R. Bagging, Boosting and C4 [C]//Proceedings of the 13th International Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1996:725-730

[42] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1):119-139

[43] Fan Wei, Stolfo S J, Zhang Junxin, et al. AdaCost: misclassification cost-sensitive boosting[C]// Proceedings of the Sixteenth International Conference on Machine Learning. San Mateo, USA, 1999:99-105

[44] Chawla N V, Lazarevic A, Hall L O, et al. Smoteboost: Improving prediction of the minority class in boosting [C]// Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Database. Dubrovnik, Croatia, 2003:107-119

[45] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140

[46] Tao Dacheng, Tang Xiaoou, Li Xuelong, et al. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1088-1099

[47] Hido S, Kashima H. Roughly Balanced Bagging for Imbalanced Data[C]// Proceedings of the SIAM International Conference on Data Mining. Atlanta, USA, 2008:143-152

[48] 毕华,梁洪力,王珏. 重采样方法与机器学习[J]. 计算机学报, 2009,32(5):862-877

[49] Domingos P. MetaCost: A general method for making classifiers cost-sensitive[C]// Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, USA: ACM Press, 1999:155-164

[50] Zheng Zhaohui, Wu Xiaoyun, Srihar R. Feature selection for text categorization on imbalanced data[J]. SIGKDD Explorations, 2004, 6(1):80-89

(上接第 22 页)

[66] Bertoni A, Folgieri R, Valentini G. Bio-molecular cancer prediction with random subspace ensembles of support vector machines[J]. Neurocomputing, 2005, 63(1):535-539

[67] Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006, 7:3

[68] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification[J]. BMC Bioinformatics, 2008, 9:319

[69] Zhou Z H, Wu J X, Tang W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1): 239-263

[70] Peng Y H. A novel ensemble machine learning for robust microarray data classification[J]. Computers in Biology and Medicine, 2006, 36(6):553-573

[71] Kim K J, Cho S B. An Evolutionary Algorithm Approach to Optimal Ensemble Classifiers for DNA Microarray Data Analysis [J]. IEEE Trans on Evolutionary Computation, 2008, 12(3): 377-388

[72] Yeang C H, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types[J]. Bioinformatics, 2001, 17(1): 316-322

[73] Lee Y, Lee C. Classification of multiple cancer types by multicategory support vector machines using gene expression data[J]. Bioinformatics, 2003, 19(9):1132-1139

[74] Berrar D, Bradbury I, Dubitzky W. Instance-based concept learning from multiclass DNA microarray data[J]. BMC Bioinformatics, 2006, 7:73

[75] Shen L, Tan E C. Reducing multiclass cancer classification to binary by output coding and SVM[J]. Computational Biology and Chemistry, 2005, 30(1):63-71

[76] Statnikov A, Aliferis C F, Tsamardinos I, et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis[J]. Bioinformatics, 2005, 21(5):631-643

[77] Braga-Neto U M, Dougherty E R. Is cross-validation valid for small-sample microarray classification? [J]. Bioinformatics, 2004, 20(3):374-380

[78] Fu W J, Carroll R J, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation[J]. Bioinformatics, 2005, 21(9):1979-1986

[79] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection[J]. BMC Bioinformatics, 2006, 7:91

[80] Wood I A, Visscher P M, Mengersen K L. Classification based upon gene expression data: bias and precision of error rates[J]. Bioinformatics, 2007, 23(11):1363-1370

[81] Dougherty E R, Hua J P, Bittner M L. Validation of Computational Methods in Genomics[J]. Current Genomics, 2007, 8(1): 1-19

[82] Hua J P, Xiong Z X, Lowey J, et al. Optimal number of features as a function of sample size for various classification rules[J]. Bioinformatics, 2005, 21(8):1509-1515

[83] Jiang H Y, Deng Y P, Chen H S, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes[J]. BMC Bioinformatics, 2004, 5:81

[84] Yoon Y, Lee J, Park S, et al. Direct integration of microarrays for selecting informative gene and phenotype classification[J]. Information science, 2008, 178(1/2):88-105

[85] Vogiatzis D, Tsapatsoulis N. Active learning for microarray data [J]. International Journal of Approximate Reasoning, 2008, 47(1):85-96