生物信息学中基因芯片的特征选择技术综述

周 昉 何洁月

（东南大学计算机科学与工程学院 南京 210096）

摘 要 随着生物信息学这门新兴学科的兴起，基因芯片技术的研究已经受到越来越多研究者的重视。目前，人们对疾病的分类和诊断的水平已经有了进一步的提高，基于基因芯片的特征选择技术在其中起到了关键性的作用。本文主要对当前基于基因芯片的特征选择技术的研究现状和各种技术方法等进行了综述，并分别从特征基因的选择数、时间复杂度和分类正确率等方面对各个方法进行了分析比较，展望了特征选择技术在基因芯片研究中的未来研究方向。

关键词 生物信息学，特征选择，基因芯片

Survey of the Gene Selection Technologies Based on Microarray in Bioinformatics

ZHOU Fang HE Jie-Yue

(School of Computer Science and Engineering, Southeast University, Nanjing 210096)

Abstract The research on the microarry technology has been paid more and more attention by the increasing number of the researchers with the advancement of bioinformatics which is a novel subject. Nowadays, the understanding of classification and diagnosis of the disease have made a great improvement for the gene selection technologies based on microarray data have played a central role in it. In this paper, the survey has been mainly made on the up-to-date research work and the methods of gene selection technologies based on microarray data. Meanwhile, analysis and comparison have been made on the respective methods from several facets, such as the number of the informative gene selected, temporal complexity, classification accuracy, and so on. Moreover, this paper also describes the future research prospect in the gene selection technology involved in microarray research.

Keywords Bioinformatics, Gene selection, Microarray

1 引言

生物科学与计算机科学是目前发展最迅速的两大学科，而作为这两大学科的交叉产物——生物信息学在基因组研究中发挥了重要的作用，基因芯片是其中的一个崭新的研究领域。基因芯片即 DNA 芯片或 DNA 微阵列，是在 20 世纪 90 年代中期发展出来的高科技产物，大小如指甲盖一般，每个芯片的基面上都可以划分出数万至数百万个小区，在指定的小区内，可固定大量具有特定功能、长约 20 个碱基序列的核酸分子。它可以大规模并行检测成千上万个基因的表达量，是目前一种新兴的生物学技术，为基因功能的研究提供了一种强有力的工具，对癌症等疾病的分类、诊断与病例学研究有非常重要的实际意义。目前在分子水平上对生物组织的研究已经取得了显著的成果，预示着基因芯片阵列的研究将可能对有效地进行疾病诊断起着关键作用。

基于基因芯片数据对疾病进行分类诊断是生物医学中重要的应用领域。1999 年，Golub 等人的研究证明，肿瘤亚型之间在基因表达上的差异可以通过对一组特异基因的表达检测进行临床诊断，并指导治疗方案的制定。随后，很多研究组开展了用 DNA 微阵列检测基因表达，用于肿瘤诊断的研究。在基于基因芯片进行数据研究分析的过程中，特征基因的选择并非一件容易的事，而且其选择的好坏对疾病的诊断起关键性作用。但是，基于基因芯片的特征选择面临很多的困难和挑战，其中最大的一个问题是：由于实验成本较高，基因样本数量常常很少，只有几十或者一两百例，而检测的基因数目相对而言很大，往往高达几千甚至几万，这其中包含大量的对于区分疾病类型无关的基因，这对研究提出了挑战。另一个问题就是这些研究数据经常包含“技术的”和“生物学的”两方面噪音数据。技术方面的噪音数据来自于很多不同的阶段，比如在基因芯片的产生过程中、在基因样本的准备过程中，等等。而生物学方面的噪音数据来自于样本的不同的基因遗传背景，或者是由于样本本身混合了杂质。本文着重对第一个问题——“高维小样本”进行讨论研究。

由于“高维小样本”的特点，常用的数据挖掘中的许多分类器对训练数据样本有较高的分类正确率，但是它们对未见过的测试数据样本表现的诊断正确率有可能很差。由于有些基因在功能上具有相似性，还有些基因对于区分疾病类型无关或者所起的作用微乎甚微，因此为了解决高维小样本的问题，对特征维数考虑进行压缩。文[53]中指出，在基因数据分析中，其特征基因的选择方法往往比分类器的选择更重要。一个基因来自两个不同类别的样本（如正常样本和疾病样本），其表达值可能不同。如果某基因在不同类别的样本中的表达值有明显区别，那么该基因就很可能对疾病的诊断预测有很强的鉴别力。如何利用这种具有高维、高相关、高噪音、数量有限的基因芯片数据，识别出对疾病有鉴别意义的特征基因组，这对数据挖掘研究提出了新的课题，并成为目前基因表达数据处理和分析的热点研究问题。

2 特征选择内容描述

特征选择算法是属于以减少特征空间维数为目标的算

周 昉 硕士生，主要研究方向为数据挖掘、生物信息学；何洁月 副教授，研究方向为数据库技术、数据挖掘、生物信息学、信息集成等。

法。文[1]指出，如果用于构造分类器的训练样本数目相对于属性的数目较少，那么包含过多的属性可能会降低分类器的性能。因此，对于高维数据，很有必要进行属性空间维数的压缩。降低维数的算法可以分为两类。第一类是特征抽取：通过转换高维属性空间数据来重新建立新的低维属性空间数据。典型的特征抽取算法包含一些线性变换算法：PCA(主成分分析)、SVD(单值分解)[2]。由于是直接作用在原始的属性空间上来产生一些新的属性，这些新的属性将不包含原有属性的具体含义，因此很难识别其含义，对疾病诊断没有任何现实指导意义。第二类是特征选择：从原有的属性空间中选择重要的特征属性，也就是剔除一些没有用的属性。一般有两种：一种是与分类毫无关系的无关属性，还有一种是与其他属性表现性能相似的冗余属性。由于此算法将保留原属性的具体含义，易理解识别，能加以应用，对现实生活也有一定的指导意义，因此广泛运用于基因芯片数据集的属性空间压缩中。为此，本文侧重讨论特征选择的研究技术。

文[37]指出一个好的基于基因芯片的特征选择方法应具有：

- 能够包含基因间的相互作用的信息；
- 基因选择的标准应该基于基因组的表现性能而不是个别单独基因与分类的相关性；
- 所选择基因里应当包含那些对疾病鉴别或疾病分类起辅助作用的基因；
- 所选择的基因应该是与疾病紧密相连的，对鉴别不同的疾病能力强，能为研究疾病的病因提供重要的线索，而不是因为其细胞的组成结构或成分不同而被选择；
- 所使用的方法应该尽可能地合理高效，并能找到所含特征基因个数较少的典型基因组。

3 特征选择的算法

目前已有的方法可分为三类：过滤(Filter)法、缠绕(Wrapper)法和嵌入式(Embeded)法。过滤法基于数据本身的内在结构信息而不依赖于分类算法对子集的评价，适合较大的芯片数据集。缠绕法依赖于特定分类器的评价指标，将分类算法嵌入到特征选择过程中，以达到最大分类准确率为引导的特征选择算法。嵌入式法利用具有分类功能和具有特征选择功能分类器算法，在分类的过程中进行自动特征选择，即分类和特征选择并行。

3.1 过滤法(Filter)

过滤法即传统的基于单个基因的选择方法。根据单个判别标准值来对基因属性进行排序、选择的方法比较简单，广泛运用在基因表达数据分析中，并证明有效[6,8,10,12]。虽然这些算法对于排列每个基因属性重要性(即该属性在对疾病诊断过程中所起作用的程度)是采用不同的判别标准，如信息增益[3]、t统计量(t-statistic)[9]、马尔可夫毯(Markov Blanket)过滤法[11]、信噪比(signal noise ratio SNR)[40]、Relief[33,34]、基于熵(entropy-based)等，但是它们的大致流程相同。首先，基于某种评估，方法针对每个基因进行单个评估以反映该基因与相关类的关联度；然后，根据每个基因的评估值进行排序，选择排序排在前列的基因。

3.1.1 基于不同判别标准的过滤法

基于SNR过滤法

Shipp等[40]采用SNR方法，在弥漫大B细胞淋巴瘤(diffuse large B-cell lymphoma DLBCL)数据上选择出30个特征基因对用于DLBCL与FL(follicular lymphoma 滤泡性淋巴瘤)进行分类，达到91%的分类正确率。

Liang Goh等人[41]提出了一种将Pearson相关系数(Pearson correlation coefficient PCC)与信噪比(SNR)相混合并结合渐进分类器(evolving classification function ECF)的方法，同样用淋巴瘤癌数据进行实验。将相关系数阈值设置为0.6时，从7129个属性中只需选择9个特征属性就可以达到100%的分类正确率。而单采用SNR对属性过滤选择，需要205个基因才能达到100%的分类正确率。而且，众所周知，无偏置的特征选择的分类正确率肯定低于有偏置的特征选择的分类正确率。该方法的无偏置特征选择的平均分类正确率为91%，平均只需10个特征基因，而Shipp需要30个特征基因才能达到91%的正确率。

基于Relief过滤法

Relief[33,34]是公认的效果较好的filter式特征评估方法，是以属性区分"相近"样本的能力作为评估属性重要性的标准，在一定程度上考虑了属性间的相关性。

文[33]先使用Relief算法删除不相关属性，然后使用K-means算法对属性进行聚类，删除冗余属性，最后是一个组合的特征选择算法，目的是删除不相关和冗余属性。但是此方法的一大缺点是不能辨别冗余特征。针对此不足，文[34]提出了基于Relief的组合式特征选择算法：ReCorre和ReSBSW。该两种算法先利用Relief过滤掉无关特征，然后采用相关分析以及顺序向后搜索(SBS)的Wrapper算法去除冗余特征。实验结果表明，ReCorre适合特征维数较多、无关特征和冗余特征较多的数据，而ReSBSW虽然分类准确率高于前者，但是运行效率低，因此在高维数据的应用上受到限制。

基于t-test过滤法

李丽等在文[28]中，运用改进的调整 p 值的t-test法和非参数评分的两种过滤特征基因选择算法。利用支持向量机，从分类器性能、支持向量吻合度、错分样本吻合度和训练样本逐步增加时SVM的稳定性四个方面，对用该两种过滤法所选择的特征基因子集与未进行选择时基因集进行比较。结果发现，调整 p 值的 t 检验和非参数评分具有一定的有效性，而且调整 p 值的 t 检验尤其明显。

基于马尔可夫毯过滤法

Xing等人[11]在2001年将Markov Blanket(Pearl,1988)过滤法应用到基因芯片数据分析中。T. A. Knijnenburg等在文[39]中提出基于离散化的特征属性运用Markov Blanket过滤法对小样本数据进行特征选择将产生不满意的结果，导致我们对这一方法的有效性进行质疑。

基于小波变换过滤法

由于小波变换(wavelet transform)技术能进行多解分析和空间频率定位，因此该技术在生物信息领域越来越受欢迎。Li Shutao等人[52]将此技术运用到基因芯片数据特征提取方面，提出了一种基于离散小波变换(Discrete Wavelet Transform DWT)的特征基因提取方法。将此方法作用在leukemia和Colon Cancer两个公共数据集上，得到的最佳的分类正确率分别是100%和93.55%，其结果优于其他的算法，除了JCFO[53]。JCFO在这两个公共数据集上的分类正确率分别是100%和96.8%，但是基于DWT的方法比JCFO计算速度要快。

3.1.2 根据不同需要改进的过滤法

有效删除冗余基因

基于某种评估标准，根据基因与相关类的关联度排序从而选择评估值高的作为被选基因，存在一定的问题，因为在同一代谢通路上的基因，其功能相关的基因表达倾向于高度相关。如果一个基因其评估值高，那么与它密切关联的其他基因同样具有高的评估值，因此存在冗余的问题，而该问题的存在可能会使分类结果发生偏移进而导致错误的结果。针对此问题，Y. Wang 在文[38]提出了一个将基因评估排序与聚类分析的思想相结合的新混合方法，目的是选择一组非冗余的具有高鉴别能力的特征基因并且与分类任务密切相关。该方法的思想是首先基于某种过滤标准对原始基因组进行处理，排序选择一个数目相对于原来较小的具有高评估值的基因组，然后将分级群聚(hierarchical clustering)运用到所选出的子基因组中，产生一个系统树图，最后对该系统树图进行分析，选择出特征基因。与只基于评估标准来选择特征基因的方法相比，该算法能选出尽量少的基因属性，同时其分类的正确率相对较高，能仅用 5 个基因在 ALL/AML leukemia 数据集上达到 100%的正确率，用 3 个基因在 colon tomor 数据集上达到 91%的正确率，使用 26 个基因在 MLL leukemia 数据集上达到 100%的正确率。

有效处理样本数在不同类之间的不平衡

文[37]指出过滤法存在的一个不足是基于该类方法所选择的特征基因可能没有较小的类内偏差。针对此不足，Cho 等人[42]在 2003 年提出了一个新的衡量各个基因的相关度的判别标准，即用样本到类质心距离的平均值和标准差来选择特征基因。该方法不仅适合两类样本分类，同时适合多类别样本分类。同时通过对度量进行加权以解决样本在各个类中的数目不同而影响统计结果的问题，在没有降低分类正确率的前提下，降低了时间复杂度。分别在 leukemia(包含 AML、ALL 两类)和 small round blue cell tumor(SRBCT，共有 BL，EWS，RMS，NB 四类)两个数据集上进行实验，在 leukemia 上达到最小测试误差为 4.06%，在 SRBCT 上的最小测试误差为 0.96%。

考虑到一些方法(特别是非参数选择法)都没有考虑不同类别中样本大小的不平衡问题，李建中等在文[29]提出了一个新的与数据分布模型无关的基因选择方法。把基因在属于相同类别的不同样本中表达值之间的大小差距叫做类内差距，而在不同类别的不同样本中的表达值之间的大小差距称为类间差距。一个理想的具有高鉴别能力的基因，它的类间差别应该较大，同时它的类内变化要较小。综合考虑基因的类间差别和类内变化，该文提出了新的计算基因鉴别能力的方法。在两个著名的数据集 leukemia 和 SRBCT 上的实验证明此方法能达到较好的分类精度：对经预处理后的 SRBCT 数据，从 2308 个基因中选出 74 个特征基因，训练精度和测试精度都是 100%，而 Cho 等人[42]选出的 21 个基因，获得的平均最小测试误差为 0.96%；对经预处理后的 leukemia 数据，从 3571 个基因中选出 30 个特征基因，训练和测试精度到达 100%，而 Cho 等人[42]得到的平均交叉验证误差为 4.06%。用该方法所进行的基因选择较稳定，不会受到当样本数目出现偏斜时对基因选择结果的影响，而且该方法可以直接应用于多种类别的问题。

有效处理属性间的相互关系

一般的过滤法都忽略了基因之间的复杂关系，因而容易导致冗余基因的存在。两个所谓鉴别能力高的基因组合在一起并不能保证就能产生一个更好的分类器。

为了同时考虑特征属性对分类所起的作用和特征属性之间的相互关系，文[21]引入了一个新的平衡的信息增益的方法来评估每个属性的作用。实验证明，这种方法对特征基因的选择具有有效性和鲁棒性。首先，对所有属性计算其平衡的信息增益值，删除一些没有超过所设标准值的属性，然后在剩余属性中选出所需的特征子集。递归循环从剩余属性集合中选出其平衡信息增益值最大的特征属性，移出剩余属性集合，放入被选属性的子集中，然后计算剩余属性集合中所有属性与被选属性集合的重要性。如果其值低于所设置的标准值，则直接从剩余属性集合中删除，直至选择出所需数量的特征属性，循环结束。

Xu Xian 等人在文[43]中提出了 BSFF(Boost Feature Subset Selection)算法，采用 bootstrap 技术来提高基于对单个基因评估排序的过滤法的性能。此方法特征基因是从 bootstrap 样本集 S^M 而不是原始的训练样本集中选出，S^M 是一个多重集，集合中样本是基于依据已被选择的特征基因在不同的 bootstrap 样本上的表现性能而动态改变的概率 $p(s)$ 进行随机选择，其集合容量大小一般定为原样本集合的两倍大。算法中还引入了样本最差集合(the worst set of sample)概念，即针对某个基因 g 对样本集($S^M - s$)集合中的每个样本 s 进行评估值计算，依据评估值大小进行排序。评估值高的，其相应的被排除的样本 s 即被视为最差的样本。该算法主要是为了改进过滤法的性能，因此采用了两个简单的、普遍使用的算法作为测试：t-score 和 SNR。将其融入 BFSS 算法中，分别产生了 boost t-score 和 boost SNR 两个新算法。将这两个新算法分别作用在三个基因芯片数据集上：Colon Cancer、Leukemia 和 multi-class cancer。实验结果证明，基于 boost 算法的过滤法能产生更高的分类准确性。在 colon cancer 数据集上，其性能改变得最明显，能提高 4%之多；对于 leukemia 数据集，原始的平均分类正确率已经达到 95%，因此提高的空间就很有限；对于多类数据集，更有理由希望所选中的特征基因能反映不同类的特征，因此采用 boost 算法，所有分类正确率都提高了。

有效转换空间的维数

从具有超高维空间超小样本的基因芯片数据中进行特征选择是一个组合优化问题。随着特征集合的增大，搜索的组合空间也随之爆炸增长，这是一个 NP 完全问题。

张军英等人从另一个角度思考该问题，将基因空间的基因选择问题转化为在病例空间中的数据分析问题，在文[30]提出了基于类别空间的基因选择。空间的维数仅为基因空间中样本的类别数，空间中的样本数是基因空间的维数(即基因的总个数)。运用主分量分析(PCA)投影获得一个计算基因对分类贡献的基准，以此基准分别计算每个基因对分类的贡献并加以排序，最终实现对基因的选择。该方法与其他基因选择方法如 SNR 方法进行实验比较，发现本方法所选择的基因使数据的可分性提高，同时也得出基因在数据可分上不是单独起作用的，而是集体作用的结果。因此，在定义基因在不同病类上表达数据的变换情况时是以基因数据的总体表达情况为参考基准，每一个基因对分类的贡献计算又是单个基因独立完成的。

封举富等人[32]也考虑利用 Fisher 线性判别的原理将算法的计算规模从依赖特征个数转向依赖于样本的个数。Fisher 线性判别模型的基本思想是把样本投影到一条直线上，使得在这个方向上的直线上样本的投影能分得最好。文

[32]提出了快速 Fisher 优化模型，利用再生核理论的结论，将投影向量表示成样本的线性组合，大大提高了计算速度。

有效地降低特征基因的个数

文[54]提出了一种从大量的基因空间中提取最有用的特征基因方法，其目标就是抽取出的基因能在不同类的样本之间展示出最大的鉴别能力。此方法首先计算每个基因的最大程度的纯化分类的阈值，然后计算每个基因的鉴别能力。根据每个基因的鉴别能力值的大小对基因进行降序排序，选出排在最前的基因即为特征基因。将此方法作用在 Leukemia，Lung cancer，Prostate cancer 和 Diffuse large B-cell lymphoma4 个公共基因芯片数据集上，其结果是在 Leukemia 上 6 个基因能达到 97.06%的分类正确率，在 Lung cancer 上 4 基因能达到 99.33%，在 Prostate cancer 上 2 个能达到 100%，在 DLBCL 上 3 个能达到 100%。无论是在分类正确率还是其特征基因选择的个数上，结果都明显优于其他方法。

3.1.3 过滤法的特点

过滤法是根据每个属性的评估值进行选择最优的特征子集，是一个快速收敛的算法，而且计算复杂度低。但是，由于过滤法在特征评估时与分类器的决策机制相脱离，而且它忽视了属性之间的相互关系，因此虽然所选中的单个特征属性可能具有较高的鉴别疾病和分类的能力，但是将所有选择的特征属性放在一起共同作用，其性能不一定是最好的。而且，该类方法也会将一些在鉴别疾病和分类上对其它特征基因起到辅助支撑作用而自身的鉴别能力不高的基因忽略，所以由过滤法选择的特征属性子集并不一定是最佳的。

3.2 缠绕法(Wrapper)

缠绕法[7]是将分类算法嵌入到特征选择过程中去，将分类器的输出作为一个选择标准，从而保证在每次循环过程中，选择出的特征子集将比先前选择出的特征子集具有更好的性能。

缠绕法是将特征选择和分类器相结合的方法，而分类器在机器学习方法中，分无监督学习和有监督学习，因此缠绕法又可分无监督学习缠绕法和有监督学习缠绕法。

3.2.1 无监督学习

基于无监督的特征选择方法是一种在缺乏先验知识即样本类别未知的情况下进行特征基因挖掘的方法。运用无监督的聚类，目的就是把具有相似性的对象聚到一起，而基于无监督的特征选择方法就是把对样本分类起作用的基因聚集在一起，起到降低维度的效果。目前常用的方法有分层聚类法、k-均值聚类[4,5]、自组织映射图网络。

文[4]依据特征对分类的影响和特征之间相关性，利用 k-均值聚类算法进行特征选择。由于缺乏先验知识，因此 k 并不能事先确定。文[5]引入了模式质量的概念，有效地评价基因在疾病分类中的鉴别能力，使 k 作为一个变量来处理，从 2 开始递增，对每一个 k 值都根据模式质量对每个特征基因进行评估，选出最优的基因集作为特征基因集。k 值递增重复进行，直到找到一个满意的 k 值使得对应的模式质量最大。

虽然基于无监督学习的特征选择方法对生物学研究有一定的意义，但是毕竟缺乏样本的表型信息，精确度不够高。

3.2.2 有监督学习

在有监督的学习中，特征选择算法是在假设已知一些或全部基因数据的额外信息(如这些信息的功能分类、已知是正常样本或者疾病样本等)前提下，从众多的属性中寻求一组最优的特征属性子集，最大限度地提高模式分类的准确率。

在有监督的学习中，常常需要建立一个分类器，该分类器可以通过样本的表达数据来预测样本的类型。在基因芯片数据分析中，常常是在已有数据的基础上建立分类器，利用所建立的分类器对未知样本的功能或者状态进行预测。但是数据的维数过大常常会影响分类器的分类性能，所以必须先进行特征选择，降低维数。分类器在特征选择中的作用就是评估选择的特征子集的性能。常用的分类器主要有：支持向量机(SVM)、决策树(Decision tree)、人工神经网络(ANN)等。构建训练集的方法有多种，传统的有三种[14]。

1)Bagging：在原训练集上采用有放回抽样，每次随即抽取与原训练集等同大小的集合，其特点是每一副本训练集包含原训练集的 63.2%，由该副本作为训练集，余下的样本作为实验集；

2)采用无放回随机抽样，每次抽取样本集的 $1/n$ 作为实验集，余下的样本集作为训练集；

3)n-倍交叉验证(leave-one-out cross validation)，将样本集分为 n 份，选取其中一份作为实验集，余下的 $n-1$ 份作为训练集。如此循环 n 次，产生不重叠的训练集和实验集。

大量实验证明，支持向量机较适合处理基因芯片数据这种样本少、维数高的数据集分类和特征选取问题。因此，在本文中讨论的分类器主要是支持向量机。

3.2.2.1 基于神经网络的缠绕法

文[18]是以基因对 BP 神经网络输出函数的灵敏度为依据，反向递归去除灵敏度小的若干基因，得到一组嵌套的候选特征基因子集。然后以支持向量机为分类工具，从中选择错分率最低的候选基因子集。候选基因子集的形成过程是：训练 BP 网络模型，根据网络的结构参数计算基因对 BP 输出的灵敏度，去除灵敏度最小的若干基因，剩余的基因组成一个候选特征基因子集，然后用此候选基因子集重新训练 BP 网络模型，刷新基因灵敏度，形成维数更小的候选基因子集，直至子集为空。

3.2.2.2 过滤法和缠绕法相混合的方法

基于 SNR 的缠绕法

文[17]结合了 Filter 和 Wrapper 方法，修正了 Golub 等人提出的“信噪比”指标，以全面衡量基因所蕴含的样本分类信息，并以此为依据滤除与分类无关的基因。在进行分类无关基因滤除时借鉴了 Filter 方法，在冗余去除时采用基于 Wrapper 的方法。先计算每个基因的分类信息指数，过滤分类无关的基因，对基因分类能力进行检验，然后计算任意两个基因表达水平间的相关系数，进行冗余基因的排除，最后基于 SVM 分类模型进行灵敏度分析。针对分类特征的分类模型而言，通过依次去除对决策影响最小的分类特征，每去除一个分类特征后，都将得到一个新集合，用此集合训练 SVM 分类模型，得到新的决策函数，导致必须根据新得到的决策函数重新计算各剩余分类特征的灵敏度的值，以找到具有最佳的分类能力且所含特征最少的特征子集作为最终的分类特征集合。该方法对于冗余基因的剔除有一定的实用性。

基于 Relief 的缠绕法

文[35]借鉴递归特征筛选(Recursive Feature Elimination，RFE)，基于 Relief 算法，提出了 RFE_Relief 算法进行特征基因的选择。然后结合分类模型，采用了 SVM 分类模型的灵敏度分析方法，分析各个特征影响模型输出能力的大小，进行冗余基因的排除。经实验，此方法所提取的特征集合与采用“信噪比”、基于 t-test 的基因排序法和 Relief 方法所提取

的属性集合相比，所含的分类特征最小，而且分类能力最强，但是计算复杂性较大。

基于 t-test 的缠绕法

Chu Feng 等人[47]首先用 t-test 方法选出一些重要的基因，然后运用 FNN(fuzzy neural network)递归循环作用在训练集和验证集上，结果选择出较少的基因而且达到了较高的分类正确率。得出结论：FNN 分类器不仅帮助研究者来区别用常规的方法很难区分的疾病，而且还能提取较少的与疾病密切相关的重要的特征基因。此方法作用在两个公共的基因表达数据集 Lymphoma 和 SRBCT 上，产生了明显的效果。在 Lymphoma 上，Tibshirani[49]提取了 48 个基因，达到了 100%的分类正确率，而 Chu Feng 仅用 5 个基因便达到了 100%的分类正确率；在 SRBCT 上，Tibshirani[48]提取了 43 个基因，达到了 100%的正确率；Deutsch[50]提取了 12 个基因，获得了 100%的分类正确率，而 Chu Feng 仅用 8 个特征基因便获得了 100%的分类正确率。

基于 spectral biclustering 的缠绕法

以往的基因选择方法虽然能达到 100%的分类正确率，但是其所选择的特征基因的个数太大。如此多的特征基因使得很难分辨到底哪几个基因与疾病密切相关，同时特征基因多了也增加了样本诊断的代价。因此，Liu Bing 等人[46]借鉴了 Kluger[45]的 spectual biclustering 的思想，提出了一种新的、有效的半监督基因选择方法来尽可能选择个数少的特征基因。Kluger 首先提出了用 spectual biclustering 来处理基因芯片数据，但是并未运用到基因选择上。该基因选择方法使用 spectral biclustering 方法获得了 s 个最优的分类本征向量 $\vec{v}_1,\vec{v}_2,\cdots,\vec{v}_s$，用这 s 个本征向量对基因进行排列和选择。首先计算每个基因变量与每个本征向量的相关度(相似度)：$R_{i,j}=\dfrac{(\vec{g}_i)^T\vec{v}_j}{\|\vec{g}_i\|_2\|\vec{v}_j\|_2}$，$i=1,2,\cdots n$，$j=1,2,\cdots,s$。$R_{i,j}$ 的值越大，则说明第 i 个基因与第 j 个本征向量相关度越大。对每个本征向量，对 $R_{i,j}$ 的值的大小进行排序，选择出前 l 个基因。然后从对应每个本征向量的 l 个基因中选出 1 个基因，因此共有个 l^s 组合。对每个组合用 SVM 进行测试，其分类正确率最高的组合就是最佳特征基因组合。将此方法运用到 lymphoma 数据集和 liver cancer 数据集。lymphoma 数据集共有 3 类 CLL，FL 和 DLCL。实验结果发现：如果基因 1622X 的表达值大于－0.8 时，则该样本属于 DLCL 类；如果基因 1622X 的表达值小于－0.8，并且基因 2328X 的值大于 1.1，则该样本属于 CLL 类，其余的则属于 FL 类。对于 live cancer 数据集，选出了一个基因 CLID，达到了 98.7%的分类正确率。因此，如果基因 IMAGE 301122 的表达值小于 0.003，则医生能诊断该患者患有 HCC，否则没有患 HCC。由于该方法能在缺乏知识的情况下找到最少的特征基因，因此此方法也非常适用于那些没有类标签的数据库进行特征基因的选取。

3.2.2.3 基于集成分类器的缠绕法

实验证明[15]，集成的分类器其性能比单一的分类器的性能要好，是将一组独立的分类器在某种程度上进行组合进而对新的样本进行分类，因此有人将此技术运用到特征选择算法中。

基于递归分类树的集成缠绕法

在生物医学上，因为在同一代谢通路上的基因其功能相关的基因表达倾向于高度相关，有时发现冗余基因对疾病的研究也是有指导意义的，而一般的特征选择算法则侧重排除高相关，冗余特征基因就无法选择出来。

文[14,16,20]能有效解决这个问题。文[14]借鉴有监督学习中的分类器集成决策技术[15]，提出了一种基于递归分类树的集成特征选择方法 EFST(Ensemble Feature Selection based on Recursive Partition-Tree)。该方法是由样本集构建不同分布结构的训练集和实验集，以类纯度指标最大和分类错误率最小为特征选择的评价标准，在训练集上反复进行训练学习。递归分层识别基于训练集上的一系列特征属性组，由若干特征属性基因组综合集成最终的特征子集。训练集的构建采用的是 n-倍交叉验证的方法。每一组训练集都采用分类树的方法，从根节点开始在特征属性空间中做一次穷尽的搜索，递归找出当前能在最大程度上降低划分类别的杂质度的特征属性及其阈值，将所在节点的集合根据该特征属性的阈值划分到左右两个节点，然后分别对左右两个节点进行如上操作，直至叶子节点中只包含了相同一类的样本，或者终结点仅包含最大允许数目的样本。该算法采用 Gini 差异性指标为节点的杂质函数：$E(t)=\Phi(p(\omega_1|t),p(\omega_2|t),\cdots,p(\omega_J|t)))=1-\sum_{j=1}^{J}p^2(\omega_j|t)$。如果所有类等同地混合在该点，则杂质函数值最大；而当该结点只包含一个类时，其杂质函数最小。对每个训练集计算后都会得到一组候选特征基因子集，对每个候选的特征基因子集运用到实验集上进行显著性检验，删除没有达到标准的特征子集，接着就对剩余的特征子集中的每个被选出的特征基因的被选择强度进行计算：$FV(g_k)=F(G_1,G_2,\cdots,G_m)=\dfrac{\sum_d w_d I(g_k,G_d)}{\sum_d w_d}$，最后保留被选择强度高于一定临界值的特征组成最终的特征子集。

EFST 算法能有效处理基因表达谱数据高相关的特点，能发现与疾病病因关联的基因信息，而且还有较强的降维能力，对支持向量机(SVM)和传统的模式分类 Fisher 判别、最邻近(NN)和 Logistic 非线性判别法有较强的适应性，但是从计算过程中看出其时间复杂度较高。

基于遗传算法的集成缠绕法

也有人将遗传算法(GA)运用到特征基因选择问题中[22~27,19]，因为可以通过每一代产生不同的个体，使得不同的基因组合能够被评估，因此进化算法比基于基因排序的特征选择算法更有优势。

文[31]即是一种基于遗传算法的集成特征选择方法，文中引入了集成特征选择的概念，其不仅需要找到关系密切的基因组，而且需要使得这组基因在不同的分类器上表现的差异增大。首先通过随机的选择不同的子集来产生一组分类器个体，然后通过使用遗传算法中的交叉和变异操作子作用在特征子集上，产生新的候选分类器。

文[19]提出了一个混合模型，是一个包括模糊逻辑(fuzzy logic)、GA 和 SVM 的基因特征选择技术。在第一步预处理中引入了模糊逻辑，通过聚类相似基因来大大减少维数；第二步基于 Wrapper 方法结合使用 GA 和 SVM 进行基因属性的选择，并将具有高性能的特征子集记录到一个文档中，以便进一步分析；最后，再次利用 GA 和 SVM 对已选出的子集对实验集进行验证，选择出最具有判别力的基因子集。该算法作用在 Leukemia 和 Colon 两个基因芯片数据集上，对于 Leukemia 数据集，需要选择 25 个特征基因才能达到 100%的正确率，而已有算法在 Leukemia 能选择更少的特征

基因达到100%的正确率;但是此方法对于Colon通过选择10个特征基因可以达到高达99.41%的正确率,比已有的算法提高了分类正确率。此方法仍可以加以改进。比如:没有对基因属性的个数加以考虑,可能运用此方法虽能找到最优的子集,但是不一定是数目最小的子集。

文[55]提出了一种新的概率模型建立遗传算法(Probabilistic Model Building Genetic Algorithm PMBGA)——随机概率模型建立遗传算法(Random Probabilistic Model Building Genetic Algorithm RPMBGA),从基因芯片中提取较高分类正确率的个数较少的特征基因组。与一般的遗传算法通过交叉和变异算子来产生新一代个体不同,PMBGA是通过计算上代的已选中的个体的概率分布取样来产生新一代的个体。每代的每个个体的基因选取是取决于该基因在该代的概率,而该基因在该代的概率不仅与其上代的概率有关,还与其边际分布有关,每个个体的适应值大小取决于在验证集上的分类正确率和该个体选择基因的个数。该算法作用在三个基因芯片数据集上,其结果优于一般算法。

对于某个特定的数据集,基于不同理论基础的不同的基因选择方法的结果也是不一样的。一种方法对于某些数据集来说能产生很好的结果,但是对于其他数据集而言表现就很差,这是因为每种方法考虑的侧重点不同。影响其结果的性能有多种因素,比如冗余基因的存在、基因间的相关作用和相关性、在选择特征基因的偏置问题以及对基因评价值的排序标准问题。因此对于一个特定的数据集,很难确定哪个是较优的特征基因子集。文[51]提出了利用遗传算法将多个基因选择方法的好的结果综合起来改进特征基因子集的选择,目的是有效利用不同特征选择方法的有用信息来选择更好(更少的特征基因的个数或更高的分类正确率)的特征基因组。将此方法作用在Colon Cancer和Prostate Cancer两个基因芯片数据集上,对于Colon Cancer而言,该方法只需选择9个特征基因就可以达到100%的分类正确率,无论是在分类正确率上还是在特征基因的个数上,其结果都优于SVM-RFE,t-test,entropy-based三种算法;对于Prostate Cancer而言,该方法能达到94.1%的正确率,其结果优于SVM-RFE和entropy-based,但是略差于t-test。该方法仍存在可以改进的地方,可能结果会更好。如果在考虑遗传算法的适应度时,不仅考虑最大化分类正确率和最小化特征子集的个数,还包括考虑基因间的相互关系;调用的算法的种类太少,可以将考虑基于相互关系的算法也考虑进来;遗传算法的群体个数和遗传迭代个数太少。

3.2.2.4 其他方法

文[44]提出了一种新的GSVM-REF(Granular Support Vector Machine-Recursive Feature Elimination)算法,将统计学习理论和粒计算理论相混合的方法在不同阶段以不同粒大小分别筛选无用的、多余的或有噪音的基因,能平衡地选择与类相关的特征基因。算法的第一阶段用预过滤的方法删除不相关的基因;第二阶段,在每次循环中,剩余的基因先用Fuzzy C-Means方法进行聚类,针对每一类用线性SVM对基因进行降序排列,留下具有高评估值的基因,剩余的基因又融合在一组进行递归分类选择;第三阶段,由于大部分不相关基因或冗余基因被删除,因此噪声基因影响效果也降低了。采用SVM-RFE方法对余下的基因进一步筛选,最后得到最有特征的基因子集。该算法运用在具有102个样本和12600的基因集的前列腺癌的数据集上,最好的分类正确率是99.71%,平均正确率是90.18%;而SVM-REF算法作用在相同的数据集上,最好分类正确率和平均分类正确率是94.12%和81.77%;SNR算法的结果分别是91.18%和79.41%。更重要的是,用此方法选择出17个特征基因,可以对前列腺癌数据集进行100%的正确分类。

邓赵红等人[36]根据癌基因表达序列数据集的特点,依据微分容量控制学习机DCCM,提出了一个新的特征提取度量标准NFEC,并在此基础上提出了一种循环迭代特征提取算法DCCFE。DCCM较SVM相比,其学习机假设函数集的选择范围更广泛,可以用任意一阶可微函数集作为待估计函数的目标函数集,而SVM必须取核函数。DCCFE是对基因表达数据集学习,计算各特征分量对微分容量的影响,对各特征分量表度等级,舍去基因表达数据中等级最低的特征分量,同时通过循环迭代的方式重复上述的工作,最终获得最佳的特征提取。实验表明,不论哪种分类器,DCCFE的测试精度都比较高,DCCM学习机也较其他的学习机更加稳定。

3.2.3 缠绕法的特点

由于选择出的特征子集与分类器的决策机制能较好地耦合,因此分类的准确率提高了。但是缠绕法计算复杂度高,而且在一个单个数据集上重复使用交叉验证,只为了寻找一个在验证集数据上表现良好的特征属性子集,这将导致算法以一个无法控制的速度增长。而且本质上,这个假设空间极大,交叉验证也会导致过度适应的后果。所以,当属性的个数开始增长时,缠绕法也很快变得不可行。

3.3 嵌入式方法

嵌入式方法本质上是缠绕法的一个延伸,其属性的筛选是在对一个特定的学习机训练的过程中进行的。一个典型方法[13]是将支持向量机(SVM)与递归特征筛选法(Recursive Feature Elimination RFE)相结合。支持向量机在此作为一个分类器,首先SVM作用在整个训练集上,然后对每个特征基因,计算移去该特征基因时SVM分类性能的变化。选择分类函数中关联权重绝对值最小的特征基因,并将其从训练集中移去,重复此过程直至训练集数据为空,最后一步一起删除的特征基因子集就是最优分类子集。虽然这样能得到一个理想的特征基因子集,但是其时间复杂度太高,因此文[13]中提出了一个次优化的算法,即在属性筛选的前期将一些无关的属性一批批地移出训练集。当属性个数下降到一两百时,再一个一个地筛选无用特征基因。在文[51]中将SVM-RFE,t-test[9],entropy-based分别作用在Colon Cancer和Prostate Cancer两个数据集上,结果显示在Colon Cancer数据集上,SVM-RFE只需选择16个基因即可以达到98.3%的分类正确率,而此时t-test和entropy-based的分类正确率分别为88.7%和64.5%;当作用在Prostate Cancer数据集上时,t-test只需选择2或4个基因就可以达到97.1%的分类正确率,而SVM-RFE需要选择32个基因才能达到94.1%,entropy-based的表现则更差。

结论与展望 如何从有限的、宝贵的基因芯片数据中获得对研究有指导意义的信息,是当前的一个研究热点;而如何从如此繁多的基因芯片属性中提取重要的、有意义的基因属性,成为生物信息学在基因组研究中的关键一步。

本文对特征选择算法进行了概述。过滤法是基于对每个基因进行评估,然后过滤掉评估值低的基因属性。选取得分高的基因属性作为特征属性子集,但是由于没有考虑到属性与属性之间的关系,虽然该方法简单快速,但是选出的特征子

集不一定是最优的，甚至可能是最不好的。而缠绕法虽然能选出最优的特征子集，但是算法太过复杂。

针对基因芯片数据这种高维、高相关、高冗余、小样本的特点进行特征的选择，目前已存在一些算法，每种算法作用在不同的基因芯片数据集上其表现的性能也不同。例如某种算法对某个基因芯片数据集能得到较高的分类正确率，而对于其他的基因芯片数据集所能得到的分类正确率的结果就不如其他特征选择算法，因此不存在一种通用的算法能很好地处理所有的基因芯片数据集，而且有些算法仍存在需要改进的地方。比如遗传算法虽是一个可以找到最优解的并行算法，但在高维空间中不能很好地收敛。由于基因芯片属性数量之大，造成搜索空间范围增大，而其采用随机的交叉和变异算子延长了搜索的时间，增加了一些不必要的计算，因此可以采用引入文[23]中的新的遗传算子来加快学习机的速度；另外，在对选出的基因属性的判别性能方面，可以结合多个方面但有偏重的综合模糊进行计算，可使得评估值更加精确可信。

而对于特定的基因芯片数据集，不同的特征选择算法的所表现的性能优势也各不相同。有的算法能得到很高的分类正确率，但是所选择的特征基因数却比较多；有的算法能得到较少的基因，但是时间复杂度又过大；而有些时间复杂度小的方法，其分类正确率却不令人满意；有些方法即使能有较高的分类正确率和较少的特征基因，但是其选择的特征基因与疾病的机理相关不大。因此，特定的基因芯片数据集，我们可以考虑多个算法，综合其优势，将其有效地结合起来，使得表现性能在各个方面都令人满意。

基因数据集的特征选择工作，主要目的是发掘导致疾病的根本原因，对医护人员能正确诊断有所帮助，对药物人员的药物研制有所指导，对患者可以减轻其检查疾病和治疗疾病的代价。因此，对特定的基因芯片数据，所要追求的目标是用最少的基因得到最大的分类正确率，同时还得兼顾较小的时空开销。同时，在今后的工作中，还需将各种特征选择方法选择出的基因子集和生物学知识库进行关联，查找它们的生物学功能，以便从分子水平来解释各种疾病的发病机理，对相应的特征选择算法的有效性进行进一步的评价。

参 考 文 献

1 Jain A K, Duin R P W, Mao Jianchang. Statistical pattern recognition: A review. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000,22(1):4～37

2 Anton H. Elementary Linear Algebra. 8th edition. Wiley, 2000

3 FDash M, Lin J, Yao J. Dimensionality reduction of unsupervised data [A]. In: Proc. 9th IEEE Int Conf. Tools with Artifical Intelligence [C], 1997. 532～539

4 张莉，孙钢，郭军. 基于无监督学习的特征选择方法. 见：2004中国控年会论文集，2004. 218～220

5 徐连彬，王亚东，李霞，等. 基于基因表达谱的疾病亚型特征基因挖掘方法. 生物信息学，2005. 69～72

6 Blum A L, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997,97:245～271

7 Kohavi R, John G H. Wrappers of feature subset selection. Artificial Intelligence, 1997, 97:273～324

8 Ben-Dor A, Bruhn L, Friedman N, et al. Tissue classification with gene expression profiles [J]. J Computer Biol, 2000, 7:559～583

9 Golub T R, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 1999, 286(5439):531～537

10 Li W, Grosse I. Gene selection criterion for discriminant microarray data analysis based on extreme value distributions. In: Proc. RECOMB,2003

11 Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data. In: Proc. 18th International Conf on Machine Learing. Morgan Kaufmann, San Francisco, CA ,2001. 601～608

12 Tusher V G, Tibshirani R, et al. Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 2001, 98(9): 5116～5121

13 Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Machine Learning, 2000, 46(1-3): 389～422

14 李霞，张田文，郭政. 一种基于递归分类树的集成特征基因选择方法. 计算机学报，2004

15 Dietterich T G. Ensemble methods in machine learing. In: Proceedings of the 1st Intenational Workshop on Multiple Classifier Systems. Roli F. ed. Lecture Notes in Computer Science. New York: Springer, 2000. 1～15

16 李霞，张田文，等. 特征基因挖掘的决策森林方法. 哈尔滨工业大学学报，2004

17 李颖新，阮晓刚. 基于支持向量机的肿瘤分类特征基因选取. 计算机研究与发展，2005. 1796～1801

18 刘全金，李颖新，朱云华，等. 基于 BP 神经网络的肿瘤特征基因选取. 计算机工程与应用，2005

19 Huerta E B, Duval B, Hao Jin-Kao. A hybrid GA /SVM approach for gene selection and classification of microarray data. Springer-Verlag Berlin Heidelberg, 2006

20 李霞，张田文，李丽，等. 决策树特征基因选择方法对 SVM 有效性研究. 中国生物医学工程学报，2004

21 Wu Y, Zhang A. Feature selection for classifying high-dimensional numerical data. In: IEEE Conference on Computer Vision and Pattern Recognition 2004, 2004, 2:251～258

22 David W. Opitz. Feature Selection for Ensembles. American Association for Artificial Intelligence, 1999

23 Salcedo-Sanz S, Prez-Cruz F, Campsand G, et al. Enhancing genetic feature selection through restricted search and Walsh analysis. IEEE Transactions on Systems, Man and Cybernetics, Part C, 2004, 34:398～406

24 Liu J, Iba H. Selecting informative genes using a multiobjective evolutionary algorithm. In: Proc. of Congress on Evolutionary Computation, 2002. 297～302

25 Vose M D, Wright A H. The simple genetic algorithm and the Walsh transform: part I, theory. Evolutionary Computation, 1998

26 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. Bioinformatics, 2001, 17(12): 1131～1142

27 Li L, et al. Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method. Comb Chem High Throughput Screen, 2001

28 李丽，李霞，郭政，等. 两种过滤特征基因选择算法的有效性研究. 生命科学研究，2003

29 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. 软件学报，2006

30 张军英，Wang T J，Khan J，et al. 基于类别空间的基因选择. 中国科学，2003

31 Opitz D W. Feature Selection for Ensembles American Association for Artificial Intelligence, 1999

32 封举富，时建新. 基因选择的快速 Fisher 优化模型. 北京大学学报，2005

33 Jose B, Draper B A. Feature Selection from Huge Feature Sets. In: Proc. of the 8th IEEE Conf on Computer Vision and Pattern Recognition, Vol 2 [C]. 2001

34 张丽新，等. 基于 Relief 的组合式特征选择. 复旦学报(自然科学版)，2004

35 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. 计算机学报，2006

36 邓赵红，王士同，胡德文. 适于癌基因表达数据集的新特征提取标准 NFEC 及其分类新算法研究. 生物信息学，2004

37 Lu Ying, Han Jiawei. Cancer Classification Using Gene Expression Data. Information Systems, 2003

38 Wang Y, et al. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expres-

sion data. Bioinformation, 2005
39 Knijnenburg T A, Reinders M J T, Wessels L F A. Artifacts of Markov blanket filtering based on discretized features in small sample size applications. Pattern Recognition Letters, 2006
40 Shipp M A, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine, 2002,8(1):68～74
41 Goh Liang, Song Qun, Kasabov N. A Novel Feature Selection Method to Improve Classification of Gene Expression Data. In: Proceedings of the Second Conference on Asia-Pacific Bioinformatics, vol. 29, 2004. 161～166
42 Cho J H, Lee D, Park J H, et al. New gene selection for classification of cancer subtype considering within-class variation. FEBS Letters, 2003, 551(1):3～7
43 Xu Xian, Zhang Aidong. Boost Feature Subset Selection: A New Gene Selection Algorithm for Microarray Dataset. Lecture Notes in Computer Science(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v3992 LNCS-II, Computational Science-ICCS 2006 In: 6th International Conference , Proceedings, 2006. 670～677
44 Tang Yuchun, Zhang Yan-Qing, Huang Zhen, et al. Granular SVM-RFE Gene Selection Algorithm for Reliable Prostate Cancer Clasification on Microarray Expression Data. BIBE, 2005. 290～293
45 Kluger Y, Basri R, Chang J T, et al. Spectral biclustering of microarray cancer data: co-clustering genes and conditions, Genome Res, 2003
46 Liu Bing, Wan C, Wang Lipo. An efficient semi-unsupervised gene selection method via spectral biclustering, Nanobioscience, IEEE Transations on, 2006,5(2)
47 Chu F, Xie W, Wang L P. Gene selection and cancer classification using a fuzzy neural network. In: Proc. IEEE Annu Meeting North Amer Fuzzy Information Processing Soc, vol 2, 2004. 555～559
48 Tibshirani R, Hastie T, Narashiman B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. In: Proc. Natl Acad Sci USA, Vol 99,2002. 6567～6572
49 Tibshirani R, Hastie T, Narashiman B, et al, Class prediction by nearest shrunken centroids with applications to DNA microarrays. Statistical Science,2003,18:104～117
50 Deutsch J M. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics, 2003,19:45～52
51 Tan Feng, Fu Xuezheng, Zhang Yanqing, et al. Improving feature subset selection using a genetic algorithm for microarray gene expression data. Evolutionary Computation. In: IEEE Congress, 2006. 2529～2534
52 Li Shutao, Liao Chen, Kwok J T. Wavelet-based feature extraction for microarray data classification. In: International Joint Conference on Neural Networks, Vancouver, Canada, July 2006
53 Krishnapuram B, Carin L, Hartemink A. Gene expression analysis: Joint feature selection and classifier design. In: Scholkopf B, Tsuda K, Vert J-P, eds. Kernel Methods in Computational Biology, MIT, 2004. 299～318
54 Al-Mubaid H, Ghaffari M. Identifying the most significant genes from gene expression profiles for sample classification. In: 2006IEEE International Conference on Granular Computing, 2006. 655～658
55 Paul T K, Iba H. Extraction of information genes from microarray data. In: GECCO'05, Washington, DC, USA, 2005

(上接第 128 页)

二是为每条消息设定唯一标识符(*Unique Identifier*, UID)。当搜索消息从源节点发出时,给每条消息都加上一个唯一的标识符。每个节点都需要维护一个消息的 UID 列表,当某一节点收到搜索消息后,取出消息中的 UID 字段,把它同自身所维护的 UID 列表中的值相比较。如果该 UID 已经存在于列表中,说明该节点此前已经收到过该消息,那么该节点简单地丢弃该消息;否则,节点将该消息进行转发,并将其 UID 加入到列表中。如图 8 所示,当源节点发送一条 UID 为 m3 的搜索消息给节点 a 和节点 b 时,节点 a 与节点 b 分别检查各自的 UID 列表,节点 a 的 UID 列表中没有 m3 的记录,因此节点 a 将 m3 加入到列表中,并将消息进行转发;而节点 b 的 UID 列表中已存在 m3,因此接点 b 丢弃该消息。

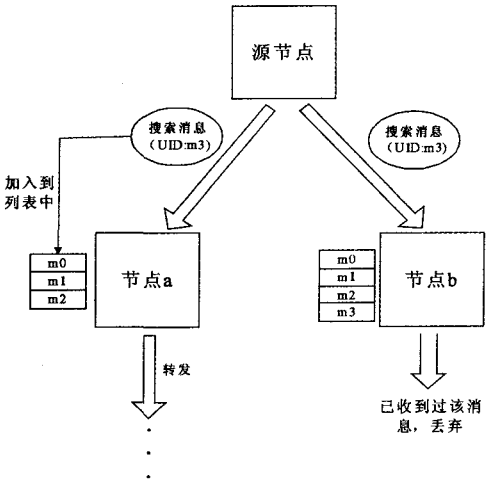

图 8 为搜索消息增加 UID

三是设定消息的重复发送次数[7]。通过事先设定消息的重复发送次数,如果搜索消息第一次发送后没有搜索到所期望的结果,则将消息的 TTL 再加一次发送。由于 TTL 值的增加,第二次搜索消息的搜索范围将比第一次搜索的范围更广。如果第二次搜索仍然未能找到所期望的结果,则将 TTL 值加 1 再发送。以此类推,直到发送次数达到预先设定的重复发送次数限制为止。

结束语 上述针对 JXTA 技术进行分析,提出了一种利用 JXTA 技术来构建 P2P 文件共享系统的设计方案,论述了对点对点实时聊天,文件共享和文件搜索三个功能模块的建模方案,并讨论了如何对传统的文件搜索算法进行改进。

参 考 文 献

1 Granville L Z, da Rosa D M, Panisson A, et al. A methodology for P2P file-sharing traffic detection [J]. Managing Computer Networks Using Peer-to-peer Technologies, 2005,7:62～68
2 Traversat B, Abdelaziz M, Doolin D, et al. Project JXTA-C: enabling a Web of things [C]. In: Proceedings of the 36th Annual Hawaii International Conference, 2003
3 Kim K, Park D. Mobile NodeID based P2P algorithm for the heterogeneous network [C]. In: Proceedings of the Second Embedded Software and Systems International Conference, 2005
4 Hsing Mei, Chang S. PP-COSE: a P2P community search scheme [C]. In: Proceedings of the Fourth Computer and Information Technology International Conference, 2004,9:416～423
5 Maibaum N, Munclt T. JXTA: a technology facilitating mobile peer-to-peer networks [J]. Mobility and Wireless Access Workshop, 2002(10):7～13
6 庄雷,潘春建,郭永强,等. Gnutella 网络的连接管理[J]. 软件学报, 2005, 16(1): 158～164
7 黄维维,黄铭钧,陈建利,等. 一种基于自配置策略的新型 Peer to Peer 平台系统[J]. 软件学报,2003,14(2): 237～246