生物信息学中的特征选择方法总结与实现

摘要

随着生物信息学的迅速发展，基因表达实验获得了大量的微阵列相关数据。然而微阵列数据具有小样本、高维度的特点，传统的模式识别技术已无法满足处理微阵列数据中大量不相关特征的要求。如何从海量的微阵列数据中高效、鲁棒地获取与疾病相关的特征基因是生物信息学领域研究的热点之一，也给模式识别、机器学习、数据挖掘等领域的研究者带来了挑战。迄今为止，特征选择技术已经被广泛地应用在癌症分类与诊断问题中。本文针对微阵列数据分析中的特征选择方法的研究现状和方法进行了阐述，编程实现了几种典型的特征选择算法，针对白血病、结肠癌数据集进行了特征选择和分类实验，并对实验结果进行了对比分析。此外，对于近年来出现的集成学习方法也进行了阐述，希望本文的工作可以为正确选择特征选择方法提供参考。

关键字： 特征选择， 微阵列， 生物信息

The Review and Realization of Feature Selection Methods in Bioinformatics

Author:Gu Yuqing
Tutor:Feng Xiaoyue

Abstract

With the rapid development of bioinformatics, gene expression experiments have obtained a large amount of microarray data. However, microarray data has a small sample size and a high dimension so that the traditional pattern recognition technology can not meet the requirement of processing a lot of irrelevant feature in microarray data. How to sort out those disease-related genes efficiently and robustly from mass microarray data has become one of the hottest research topic in bioinformatics field. Meanwhile，it has brought much challenge to researchers from pattern recognition, machine learning and data mining fields. Feature selection techniques have been widely used in cancer classification and disease diagnosis. This paper expounded the research status and a variety of different methods related to feature selection techniques in microarray data analysis. We realized several typical feature selection algorithms via programming and did feature selection and classification experiments on leukemia, colon cancer microarray datasets, and made comparisons and analysis according to the results.In addition, the paper introduced ensemble learning approach which emerged in recent years. We hope this paper can make references for selecting appropriate feature selection techniques for microarray data analysis.

Keywords: Feature Selection，Microarray，Bioinformatics

目　　录

第 1 章 绪论

1.1 研究背景及目的

随着人类基因组计划和其它测序工作的相继完成，人们已经获得了数以万计的基因序列，这为揭开生命的奥秘提供了数据基础。面对庞大的基因数据，如何利用计算机技术找出有用的信息，从而发现基因背后隐藏的生命现象是一件很有意义的研究工作。

DNA 微阵列技术可以在一次实验中同时测定成千上万个基因在不同状态下的表达数据，它的发展给生物信息学的研究提供了条件。通过机器学习的方法对基因表达数据集进行学习，将学习后的分类器用于未知样本的分类有着重大的现实意义。例如基于基因表达数据分析可以为疾病诊断提供新的手段。然而基因表达数据具有高维（成千上万）小样本（低于一百）的特点，同时数据中存在着大量的无关基因和冗余基因。这些基因的存在严重影响了分类器的性能与精度。如何从高维原始特征数据中获取与分类相关的特征子集，来对基因表达数据进行有效的降维处理改善分类效果是生物信息学领域研究的热点问题。

从上世纪 90 年代到今天，越来越多的基于微阵列数据的特征选择算法被提出。其中一些方法已经运用到肿瘤分类、癌症预测等实际问题中并且取得了显著的效果。

本文主要研究基于微阵列数据的特征选择问题，通过总结生物信息学领域各种特征选择算法对基因微阵列数据进行处理与分析，比较这些方法的优缺点，为以后遇到实际问题正确选用微阵列数据分析中特征选择方法提供参考。

1.2 微阵列数据概述

DNA 微阵列也被称为基因芯片，它借用了计算机芯片高度集成的特点，将核酸密集有序地排列在预先设定的区域内，形成微型的检测器件。微阵列数据是在一片芯片上同时测量不同样本的成千上万的基因在不同状态和不同组织中的 mRNA 表达水平而得到的一组数据，所以也称之为基因表达数据。

每一次的 DNA 微阵列实验中，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 微阵列数据集就是由在 N 次实验下所得的 M 个基因的表达值组成。微阵列数据主要为数值型并以矩阵的方式存储，可以用 $N\times M$ 维矩阵来表示。矩阵中的第 i 行、第 j 列元素所对应的数值代表在第 j 个样本中第 i 个基因的表达值。微阵列数据集是一个规模巨大的数据集合，通常涉及成千上万个基因，又因为每次的实验代价昂贵，样本的个数也有限。同时每个样本都记录了所有可测基因的表达水平，而这些基因中大部分都与样本分类无关。因此微阵列数据具有维数高、样本小、高噪声、冗余大的特点。

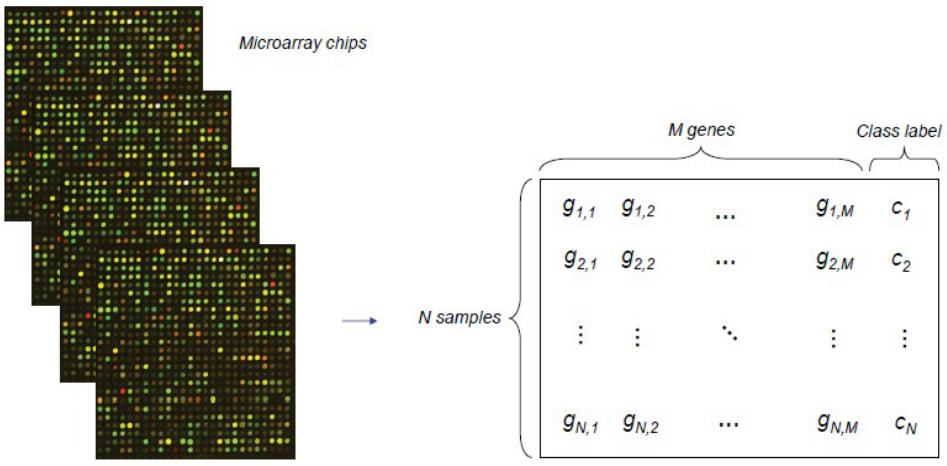

图 1-1 微阵列数据矩阵

根据特征与类别的相关性和特征之间的相关性定义，我们可以将基因

分为以下三类：相关基因、无关基因、冗余基因。相关基因是指与分类相关的基因，如果剔除将会影响分类准确率。无关基因是指对分类没有帮助的基因，它的存在与否对样本分类没有影响。冗余基因是指那些与分类相关的基因但是基因之间存在着很大的相关性，多个基因提供的分类信息不如其中一个基因提供的信息多，这些基因会对样本分类产生不良的效果。在微阵列数据集中大多数基因都是无关基因，也存在着冗余基因。所以要使用微阵列数据对癌症进行分类处理就必须要剔除无关基因、保留相关基因、消除冗余基因。

1.3 特征获取概述

对于微阵列高维数据的特征获取，常用的两类方法是特征选择和特征提取。

特征提取是指使用映射等变换方法将原有特征空间进行某种形式的变换，以得到新的特征。主成分分析是这类方法中最著名的算法，该算法对许多学习任务都可以较好降维，但对特征的理解性差。由于特征变换的新特征通常由全部初始特征变换得到，从数据收集角度看，并没有减少工作量。生物信息学中特征提取的其它方法包括小波变换、独立分量分析等。然而这些方法所提取的特征是通过某种形式变化得到的新特征属性，无法从生物学和医学角度进行解释，因此并不实用。

特征选择是指从一组原始特征中选取最有代表性的特征，以达到降低特征空间维数的目的。特征选择不改变变量的原始表达，从原始特征集中选取使某种评估标准最优的特征子集。通过特征选择，一些和分类无关或者冗余的特征被删除，简化的数据集常常会得到更精确的模型，也更容易理解。特征选择可以分为有监督学习和无监督学习两个方面。大部分的研究工作主要集中在有监督学习方面，即以分类问题为背景的特征选择。无监督学习的特征选择是个更加复杂的课题。

1.4 论文的主要工作和安排

本文以微阵列数据集作为研究的对象，根据生物信息学中特征选择算法，结合分类器对样本进行分类，目的是选择最优特征基因子集来提高分类准确率。

本论文主要包括 4 章，各章内容安排如下：

第 1 章 绪论。

第 2 章 特征选择方法。

第 3 章 集成特征选择方法

第 4 章 特征选择方法的实现

第 5 章 结论与未来展望

第 2 章　特征选择方法

特征选择的基本框架分为 4 个步骤：候选特征子集的生成(搜索策略)、评价准则、停止准则和结果验证[1]。目前对特征选择方法的研究主要集中在搜索策略和评价准则两个方面。本章从搜索策略和评价准则两个角度对几类特征选择算法进行阐述。

2.1 搜索策略

在生物信息学领域特征选择常用的搜索策略有全局最优搜索、启发式搜索和随机搜索 3 大类。

1. 全局最优搜索

全局最优搜索测量包括穷举法和分支定界法。虽然这两种方法都能得到最优特征子集，但存在着很大的局限性。在处理高维数据时算法的时间复杂度高、运行效率低，实用性不强，无法被广泛应用。

2. 启发式搜索

启发式搜索在搜索的计算量和最优性之间做了折中，它避免了大规模的完全搜索，同时也带来了丢失最优解的风险。这种方法实现过程简单快速，广泛运用在实际问题中。常用的启发式搜索有序列前向选择、序列后向选择等。在搜索过程中，特征一旦被选择或删除就不会改变，所以容易得到局部最优值。

3. 随机搜索

另一种广泛使用的搜索策略是随机搜索。此方法生成的特征子集不是顺序产生而是在随机情况下产生的。常用的随机搜索算法有遗传算法、粒子群算法、蚁群算法等。随机搜索策略能够在算法结束之前产生满足条件的特征子集，但是不能保证得到的子集是最优的。

对于一个具体的搜索算法，可能会采用多种搜索策略，例如遗传算法既是启发式搜索也是一种随机搜索算法。每种搜索策略都有各自的优劣，在实际问题中，可以根据具体环境和评价函数寻找平衡点。如果原始特征数少，可采用全局最优搜索策略；如果要求计算速度快，可采用启发式搜索策略；如果需要得到高性能的特征子集，则可使用随机搜索策略。

2.2 过滤法

以分类为背景的特征选择算法主要分为三类：过滤法、缠绕法、封装法[2]。这些分类依赖于特征选择搜索和分类模型结构的结合模式。

过滤法不依赖于分类器，通过数据的内在属性选择相关特征。该方法广泛运用于实际特征选择工作中，可以迅速对数据进行降维。过滤法的模型如图 2-1 所示。

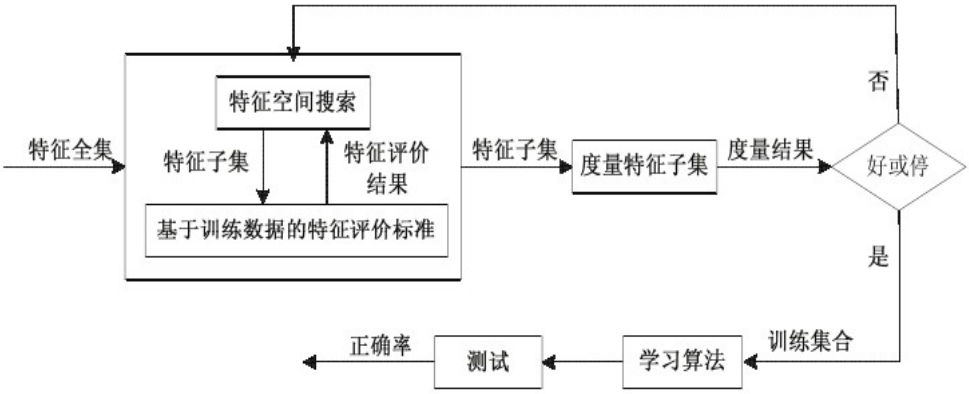

图 2-1 过滤法模型

过滤法一般使用评价准则来度量特征与类别的相关性以及特征之间的相关性。通过评价准则函数可以对候选的特征子集进行相关性评价，评价高特征子集将被选为特征基因用于后续分类。评价准则可以分为 4 类：距离度量、信息度量、依赖性度量和一致性度量[1]。

2.2.1 距离度量

距离度量通常也是分离性、差异性的度量，它包括几何距离度量和概率距离度量。针对二分类问题，特征选择的目的是找出那些使两类样本尽可能分离的特征基因。对于特征 A 和 B，如果由 A 引起的两类条件概率差异性大于 B，那么 A 对于分类的重要性大于 B。在生物信息学中使用的诸多过滤法中，Relief 评估方法及其变种 ReliefF 是基于距离度量的。距离度量的准则函数要求满足单调性或者近似单调。

2.2.2 信息度量

信息度量建立在信息论的基础上，通过衡量特征子集与类别之间的相关性和子集中特征之间的相关性作为特征子集优劣的评价标准。实际工作中通常采用信息增益(Information gain)或互信息(Mutual information)进行信息度量。信息增益是先验不确定性与期望的后验不确定性之间的差异，一般来说，特征的信息增益越大，选择这些特征训练的分类器就会有很好的性能。信息增益可以有效地选出重要特征，剔除无关特征。互信息则是描述两个随机变量之间的相互依存关系的强弱。信息度量函数是过滤法特征选择方法的核心，有多种不同的形式。下面列举几个在生物信息学中常用的基于信息度量的特征选择算法：

1. mRMR(minimal redundancy and maximal relevance)[3]

该方法从理论上证明了与最大依赖性等价，从最小冗余最大相关角度出发，使用互信息作为评价准则，具体评价函数如下：

$$J(f) = I(C, f) - \frac{1}{|S|}\sum_{s \in S} I(s, f) \qquad (2\text{-}1)$$

式中 $I(C,f)$为类别 C 与候选特征 f 之间的互信息值，$|S|$是已选特征的个数。该算法的核心是使特征子集与类别的相关性最大化，同时最小化特征间的冗余。

2. CFS (correlation-based feature selection)[4]

CFS 方法的本质是基于相关性的启发式评价函数的过滤算法。它通过评价不同特征子集的优良程度，然后选择评分最高的特征子集作为最优特征子集。其核心思想是提取与类别高度相关的特征集，同时去掉一些冗余的特征基因。CFS 特征子集评价函数如下式：

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2-2}$$

其中 M_s 是含有 k 个特征基因的特征集 S 的启发式标准，$\overline{r_{cf}}$ 是特征基因与类别之间的相关系数均值($f \in S$)，$\overline{r_{ff}}$ 是特征之间的平均相关系数。此公式是 CFS 的核心，可用来评价特征子集的预测能力和特征间的冗余程度（特征间相关度越低，特征与类别间相关度越高，评分越高）。CFS 是一种评价算法，特征子集是通过搜索算法选出来的。

基于信息度量的特征选择算法是近年的研究热点之一。其评价函数虽然有多种形式，但设计函数的核心思想都是找出与类别相关性最大的特征子集，同时子集中的特征间相关性最小。

2.2.3 依赖性度量

相关性度量使用了统计学领域的许多相关系数，这些度量方法被用来衡量特征对于类别可分的重要性程度。典型的方法有 t-test、ANOVA、Wilcoxon rank sum、BSS/WSS 方法、TNoM score 方法、Fisher 指标、信噪比(signal noise ratio SNR)等[2]。依赖性度量与信息度量相似，都是考量两个变量之间的关系，但依赖性度量一般会预先设定阈值，只有特征与类别间的关系大于这个阈值的情况下，才会认为该特征较优。两者的出发点并不相同。将依赖性度量方法应用到生物信息学中的特征选择有很多。其中 t-test 和信噪比方法的具体公式如下[5]：

1. t-test：

$$score(i) = |\mu_i(+) - \mu_i(-)| / \sqrt{\sigma_i(+)^2 / n(+) + \sigma_i(-)^2 / n(-)} \quad (2\text{-}3)$$

2. 信噪比：

$$score(i) = |\mu_i(+) - \mu_i(-)| / \sqrt{\sigma_i(+) + \sigma_i(-)} \quad (2\text{-}4)$$

其中 $n(+)$ 和 $n(-)$ 分别是两类样本的个数，$\mu_i(+)$ 和 $\mu_i(-)$ 分别是第 i 个基因在两类样本中的平均表达值，$\sigma_i(+)$ 和 $\sigma_i(-)$ 分别是第 i 个基因表达值的标准差。$score(i)$ 值越大，表示第 i 个基因对于分类越重要，将被选为特征基因。

2.2.4 一致性度量

一致性度量是一种较新的度量方式，通过考察数据集的不一致率来评价，试图找到与全集相同分类能力的最小特征子集。它具有单调、快速、去除冗余等优点，但该评价方法对噪声敏感，只适合离散数据集，而对于连续数据必须进行离散采样后才能使用。

上文分析了过滤法中的不同评价准则，选择合适的准则函数会得到较好的分类结果。过滤法是一种快速收敛的算法。它的优势在于空间计算复杂度低、速度快、容易实现，可以大量剔除无关特征。但由于对特征评估时与分类器相互脱离，并不能保证得出一个优化的特征子集，因此此方法一般用在数据预处理阶段。

2.3 缠绕法

与过滤法不同的是，缠绕法将分类算法嵌入到特征选择的过程中，根据分类精度的高低来评价特征子集的优劣，因此缠绕法的特征选择过程是依赖于分类器的。这类方法的模型如图 2-2 所示，缠绕法中用于评价特征的学习算法是没有限制的。这一方法的出现解决了过滤法无法同时考虑模

型空间搜索与特征空间搜索的问题。在缠绕法中，只要给出特征子集的搜索算法，就可以产生一系列的特征子集。但是特征子集的个数是以特征数目的指数级别增大的，因此在处理高维数据时需要采用搜索算法来寻求最优特征子集。主要的缠绕法特征选择算法有结合遗传算法和 k –最近邻方法的特征选择(GA/KNN)、遗传算法与支持向量机的结合(GA/SVM)等。

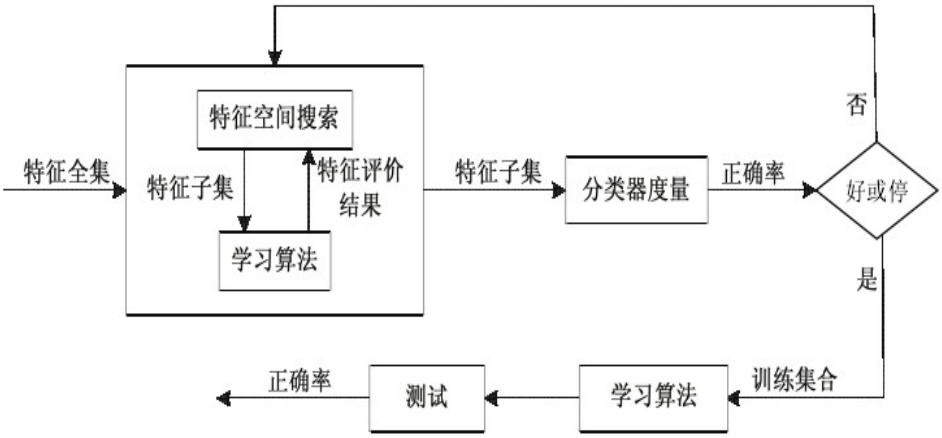

图 2-2 缠绕法模型

典型的缠绕法首先使用某种特征空间搜索策略建立特征基因子集，不同的搜索策略对特征子集的结果影响很大。通过某种搜索策略得到特征基因子集后，缠绕法将借助支持向量机、k-最近邻分类器、朴素贝叶斯等分类器直接对该特征基因子集进行分类性能评价。根据评价结果不断调整特征子集而达到探索最优子集的目的。一般算法是根据当前特征子集是否比此前选择的特征子集分率准确率更高给当前特征子集里的基因评价为 0 或 1 分，也有的根据 ROC 曲线或 LASSO 模型来评价所选择的特征子集[2]。

缠绕法采用分类算法评价选取的特征基因子集，因此具有比过滤法分类准确率高、选取特征个数少、特征冗余小的优点。但是由于反复调用分类算法会产生极高的计算量，该方法存在算法效率低、时间复杂度高、计算速度慢的弊端。另外某些搜索策略容易陷入局部最优，微阵列数据小样本高维度的特点也使得该方法存在着过拟合的风险。

2.4 嵌入法

嵌入式方法本质上是缠绕法的一个延伸，其特征的筛选是在对一个特定的学习机训练的过程中进行的。在嵌入法中，特征选择算法本身作为组成部分嵌入到学习算法里，可以看成在特征子集空间和假设空间的并集中进行搜索和缠绕法一样，嵌入法也与特定的分类算法有关。

支持向量机-递归特征剔除(SVM-RFE)是微阵列数据分析上最成功的特征选择方法之一。该方法由 Guyon 等[6]于 2002 年提出，通过使用迭代一次删除一个或多个无关基因，直到剩余基因的集合达到期望的基因数目时停止迭代，剩余的基因为算法选择的特征基因。它将每一步迭代中对 SVM 训练得到的基因决策函数作为相关的标准。SVM-RFE 算法首先在整个数据集上测试分类性能，然后计算移出每个基因后的性能变化，选择排序函数中影响最小的特征，并将其从训练集中剔除，重复此过程直到训练集数据为空，算法最后输出的是特征排序列表。SVM-RFE 算法的流程图如图 2-3 所示[7]：

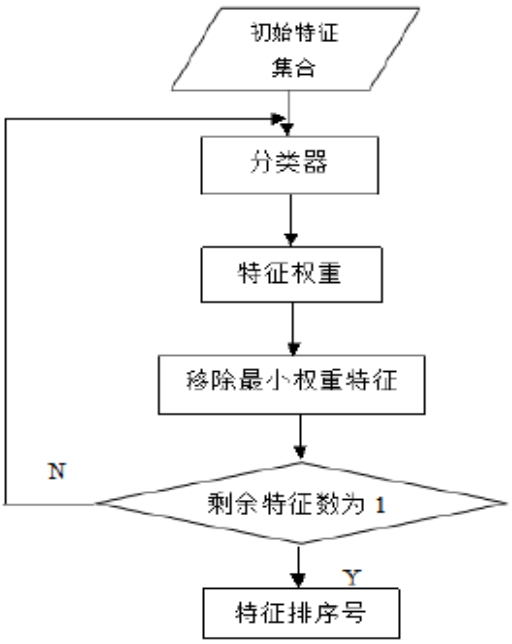

图 2-3 SVM-RFE 算法流程图

基于递归剔除的方法有很好的分类性能，此方法与分类算法相互作用，然而计算量却比缠绕法小。从某种意义上来说，嵌入法是缠绕法的扩展和延伸。

2.5 混合法

根据以上的介绍可知，每一种特征选择算法都有其优点和缺陷。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。因此使用单一的特征选择方法往往不能得到最佳效果。对于特定的微阵列数据集，我们可以考虑将多个算法结合起来，利用其各自的优势，使得表现性能在各个方面都达到令人满意的效果。

上述特征选择三种方法中，过滤法与缠绕法的混合在实际特征选择问题中运用比较多。这种混合法实质是一种两阶段组合方法，该方法先用某种过滤法选出特征然后再用某种缠绕法或嵌入法从这些特征里选出最终的特征。

为了得到一个特征基因数小且对分类贡献大的特征子集，可以使用混合法对原始特征集进行两阶段处理。首先使用过滤法剔除那些与分类无关的基因，由于过滤法计算复杂度低，可以将成千上万维的原始特征集缩小到只有几十到几百个的特征子集空间。然后在此特征子集的基础上使用缠绕法搜寻最优子集。因为过滤法已经对原始特征集进行了降维处理，后面缠绕法的计算复杂度将大大降低。目前此类方法得到了广泛的关注，研究者也提出了许多混合方法，例如先用 t-test 或信噪比进行特征基因初选，然后采用遗传算法或 SVM-RFE 进行基因精选。实验证明该方法可以通过选出很少的基因，来获得很高的分类准确率。

2.6 常用分类器

在对原始数据集进行了特征选择后，就需要使用分类器进行样本分类。在机器学习中，分类器有很多，下面主要介绍三种常用的分类器：支持向量机、k-最近邻分类法和决策树。

2.6.1 支持向量机(SVM)

支持向量机是一种有监督的学习方法，该方法能够通过分类和回归模型来分析数据及识别模式。1995 年，Cortes 和 Vapnik 提出了标准支持向量机算法。当支持向量机用于分类问题时，输入是一个用于训练的样本集和样本集对于的分类标号。之后，支持向量机使用样本集和标号训练算法建立模型。最后，输入一个没有分类标号的新样本集到这个模型中，可以得到对这个样本集中所有样本的预测分类结果。支持向量机建立在统计学理论的 VC 维理论和结构风险最小化理论的基础上，能较好地解决小样本、高维数等实际问题[9]。假定样本集合为：$\{(x_i, y_i)\}$、和分类标号：$y_i \in \{+1, -1\}$。支持向量机的决策规则如下：

$$\text{SVM}(x)=\text{sign}\ \left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x)+b\right) \tag{2-5}$$

式中，y_i 表示样本 x_i 的分类标号，求和操作是在所有样本上进行的；α_i 是分类的最大分隔边界下的拉格朗日乘子；$K(x_i, x)$ 是能够把特征空间映射到高维空间的核函数。常用的核函数主要有：线性核函数、RBF 核函数、齐次多项式核函数、非齐次多项式核函数和 Sigmod 核函数。

支持向量机的优势在于能够利用有限的样本得到最优解。这对于处理小样本微阵列数据非常适用。另外支持向量机在理论上能得到全局最优解，避免陷入局部极值问题。支持向量机与核函数的结合可以解决非线性分类问题，具有良好推广能力。因此支持向量机被广泛应用在模式识别、数据挖掘、机器学习等领域的分类问题中。

2.6.2 *k*-最近邻分类法（KNN）

k-最近邻分类法建立在通过与模式空间类比的基础上。当给定一个未知样本，*k*-最近邻算法将搜索模式空间，找出与此样本最接近的 *k* 个训练样本，这 *k* 个样本就是该未知样本的 *k* 个最近邻，未知样本将被分配到 *k* 个近邻者中类别数最多的那一类。样本之间的距离通常用欧式距离来度量。*k*-最近邻法原理简单且分类速度快，得到了广泛的应用。

2.6.3 决策树

决策树是数据挖掘中常用的分类技术，它一种用于分类和预测的树结构。其中内个内部节点表示在一个特征上的测试，每一个分支代表一个测试的输出，每个叶节点代表类分布。决策树采用贪心算法，自上而下递归地对数据进行分割处理。它首先对训练样本数据进行处理，随着树的构造，训练集被递归地分成更小的子集。决策树调用特征选择标准决定分裂节点的特征，该节点表示在一个特征上的测试。每个分支对应该特征的一个测试结果，训练集的样本根据该特征值相应的被划分到各个分支中。决策树分类过程中不需要人为设定参数并且能够处理高维数据，因此广泛用于微阵列数据样本分类问题中。

2.7 本章小结

本章首先介绍了三种生物信息学中常用的搜索策略：全局最优搜索、启发式搜索和随机搜索。然后介绍了四类特征选择算法：过滤法、缠绕法、嵌入法和混合法，并对每一类方法进行了深入研究，总结出用于生物信息学特征选择问题的典型算法。对于每一类算法，总结了其优势与劣势并作出了比较。最后介绍了三种常用的分类器：支持向量机、*k*-最近邻分类法

和决策树，用于特征选择后的样本分类。

第 3 章　集成特征选择方法

3.1　集成学习概述

面对微阵列高维数据，研究者通常首先利用特征选择算法进行降维，从而降低问题复杂度。但高维数据的降维本身就是一个很困难的问题，而且降维可能会丢失有用的特征。集成学习方法为解决高维数据的学习问题提供了一种有效的解决办法。

集成学习是近年来机器学习和模式识别领域的研究热点之一。由于集成学习方法具有处理小样本、高维度和复杂数据结构数据的独有优势而被应用到生物信息学的特征获取。集成学习通过训练多个学习系统并将其结果按一定方式进行集成，以取得比单个学习系统更高的分类准确率。如果把单个学习系统比作一个决策者，集成学习方法就是多个决策者共同进行一项策略。这种集成方法通过将多种分类模型平均结合来减少小样本训练集出现过拟合问题的可能。一些集成方法如随机森林(random forests)对处理高维数据非常有用，因为可以通过使用不同的特征子集生成多种预测模型来提高分类精度。

许多集成方法已经被应用到生物信息学中的微阵列数据分析。其中具有代表性的有 bagging、boosting 和 random forests[11]。

3.2　典型集成学习方法

集成学习方法对分类任务的改进往往通过聚集一组基本分类器作为一个强分类器来对未知数据进行预测。这也应验了“三个臭皮匠赛过诸葛亮”的俗语。一般的基本分类器可以由决策树、神经网络、贝叶斯分类器、*k*-最近邻等构成。设计这种集成方法通常会增加模型复杂度，为了使集成

方法获得更好的泛化能力一般基于经典的方差偏差理论构造模型。

3.2.1 Bagging

bagging 是 bootstrap aggregating 的缩写，其思想是对训练集有放回的抽取训练样例，从而为每一个基本分类器都构造出一个跟训练集同样大小但各不相同的训练集，从而训练出不同的基本分类器。其中训练集的获得使用的是 bootstap 算法，其核心思想和基本步骤如下：

1. 采用重抽样技术从原始样本中抽取一定数量(自己给定)的样本，此过程允许重复抽样。

2. 根据抽出的样本计算给定的统计量 T。

3. 重复上述 N 次(一般大于 1000)，得到 N 个统计量 T。

4. 计算上述 N 个统计量 T 的样本方差，得到统计量的方差。

bootstrap 是现代统计学较为流行的一种统计方法，在小样本时效果很好。通过方差的估计可以构造置信区间等，其运用范围得到进一步延伸。

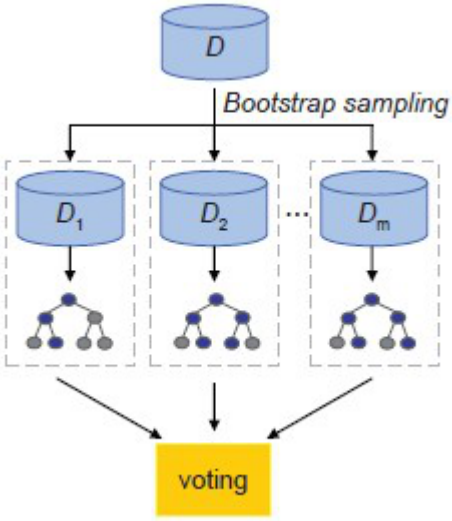

图 3-1 bagging 模型

3.2.2 Boosting

boosting 的思想是对那些容易分类错的训练实例加强学习，就好像一个人背英语单词一样，首先第一遍背完以后有一些容易记住的单词就记住了，还有一些不容易记住的，则第二遍的时候对那些不容易记住的单词多看几眼，第三遍又对第二遍还记不住的单词再多看几眼……具体实施的时候 boosting 方法是这样进行的：首先给每一个训练样例赋予相同的权重，然后训练第一个基本分类器并用它来对训练集进行测试，对于那些分类错误的测试样例提高其权重(实际算法中是降低分类正确的样例的权重)，然后用调整后的带权训练集训练第二个基本分类器，然后重复这个过程直到最后得到一个足够好的学习器。

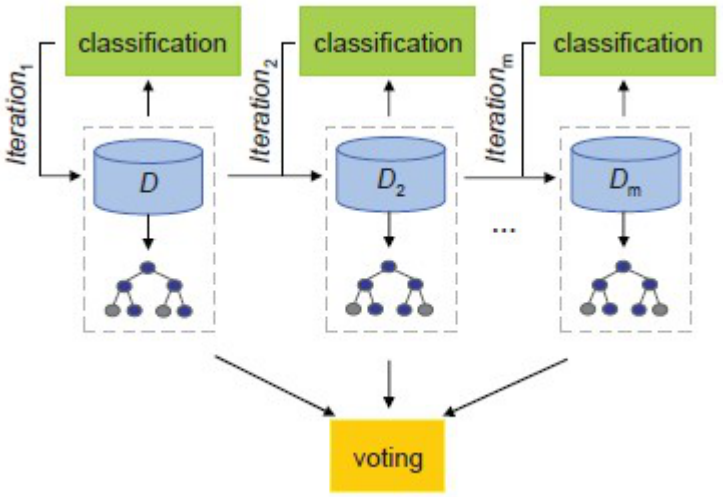

图 3-2 boosting 模型

3.2.3 Random Forest

随机森林，顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类(对于分类算法)，然后

看看哪一类被选择最多，就预测这个样本为那一类。在建立每一棵决策树的过程中，有两点需要注意 - 采样与完全分裂。首先是两个随机采样的过程，随机森林对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为 N 个，那么采样的样本也为 N 个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现过拟合。然后进行列采样，从 M 个特征中，选择 m 个($m << M$)。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。一般很多的决策树算法都一个重要的步骤 - 剪枝，但是这里不这样干，由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现过拟合。按这种算法得到的随机森林中的每一棵都是很弱的，但是大家组合起来就很厉害了。可以这样比喻随机森林算法：每一棵决策树就是一个精通于某一个窄领域的专家(因为我们从 M 个特征中选择 m 让每一棵决策树进行学习)，这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题，可以用不同的角度去看待它，最终由各个专家，投票得到结果。

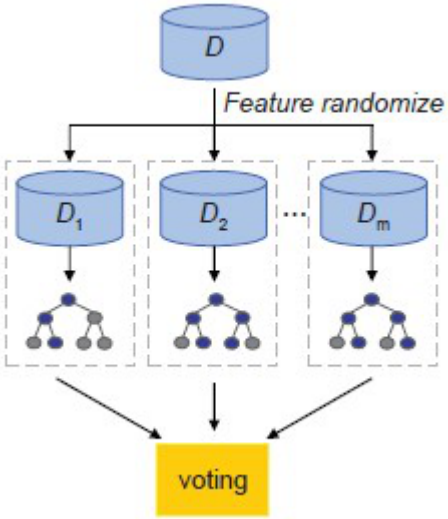

图 3-3 random forests 模型

3.3 集成特征选择方法

传统的微阵列数据处理使用一种特定的特征选择算法并将其输出作为用于分类的最终子集。借鉴集成学习的思想，不同的特征选择方法可以混合起来作为集成特征选择方法。事实证明没有哪种单一的特征选择方法是最优的，由于用作样本分类并获得相同分类准确率的特征子集可能不止一个，我们也很难决定使用哪个特征子集。因此集成特征选择方法被应用到生物信息学领域并改善特征选择方法的鲁棒性与稳定性。

集成特征选择方法是一种将各种基于排序的特征选择算法得到的特征排序列表结合生成单一排序表的方法。这种方法的实施过程分为两步。首先通过一些特征排序算法建立若干个排序列表，原始数据集中的每个特征在每个排序表中都有一个名次。然后选择某种聚集函数将第一步得到的若干排序表转化成单一排序表，每个特征在最终的排序表中有了新的次序。根据需要在最终的排序表中选择排名靠前的特征基因用于后续的分类工作。第二步使用的混合方法是集成特征选择的关键。

目前所有的集成特征选择方法在整个集成过程中除了混合方法其它都是相似的。这些集成方法使用不同的聚合函数如均值、中位数等。在均值方法中，每个特征的分数由每个特征在每个排序表中的平均值决定。而中位数方法中，每个特征混合后的分数是所有排序表中的中位数。

尽管集成特征选择方法的使用会需要额外的计算资源，但该方法在处理高维小样本问题上有着显著的效果，降低了过拟合的风险，同时也增强了特征选择的鲁棒性与稳定性。

3.4 本章小结

本章介绍了近年来出现的集成学习方法并将其应用在在生物信息学领域特征选择中。通过介绍三种典型的集成学习算法：bagging、boosting、

和 random forests，给高维小样本数据的特征选择以启发。借鉴集成学习的思想，提出了集成特征选择方法，这类方法降低了小样本数据过拟合的风险，提高分类准确率和模型的鲁棒性与稳定性。

第 4 章　特征选择方法的实现

4.1　实验数据集

用于数据分析的微阵列数据集有很多，为了验证几种典型的特征选择算法，使用 2 个微阵列数据集来进行实验。相关的属性信息如表 4-1 所示。白血病数据集(Leukemia)包括急性骨髓性白血病(AML)和急性淋巴性白血病(ALL)，该数据集中包括 72 个样本(47 个 ALL 和 25 个 AML)及 7129 个基因。结肠癌数据集(Colon)包括 40 个结肠癌样本和 22 个正常样本，每个样本有 2000 个基因。

表 4-1　实验微阵列数据集描述

数据集	样本(+/-)	训练集/测试集	基因数目
白血病	72(47/25)	38/34	7129
结肠癌	62(22/40)	31/31	2000

实验首先对数据进行标准化处理，这主要是因为后期的数据分析需要各个基因处于同一量纲之下。常用的微阵列数据标准化方法有以下两种：

1. 限制每个基因的均值为 0，方差为 1，计算公式为：

$$x_{ij}^{*} = (x_{ij} - \mu(i)) / \sigma(i) \tag{4-1}$$

式中 x_{ij}^{*} 和 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，$\mu(i)$ 和 $\sigma(i)$ 分别是第 i 个基因在所有样本上的平均值和标准差。

2. 限制每个基因的值在[0,1]范围内，计算公式为：

$$x_{ij}^{*} = (x_{ij} - x_{\min}) / (x_{\max} - x_{\min}) \tag{4-2}$$

式中 $x_{\max}$ 和 $x_{\min}$ 分别为第 i 个基因在所有样本上的最大值和最小值。

4.2 实验方法

软件：操作系统 Windows 7，编译软件 Matlab，使用 Libsvm 实现支持向量机分类器。

硬件：内存 2GB，CORE i3 CPU 2.3GHz。

实验中编程实现了 3 种过滤法：信噪比、t-test、mRMR 方法和嵌入法 SVM-RFE。使用这 4 种特征选择算法和线性 SVM 分类器对样本进行分类实验。原始数据集使用上文提到的第二种方法进行数据归一化预处理，所有基因的表达值都在[0,1]范围内。

对于分类器性能的评价使用以下两种方法：

1. 10-折交叉验证(Cross Validation)，将训练集分成 10 个子集，每个子集均做一次测试集，其余的作为训练集。10-CV 交叉验证重复 10 次，每次选择一个子集作为测试集，并将 10 次的平均交叉验证识别率作为结果。

2. 独立测试，先用训练集训练 SVM 分类器，然后用测试集评估分类性能。分类准确率为样本正确分类数与总样本数的比。

4.3 实验结果分析

4.3.1 特征基因选择结果

根据上述 4 种典型的特征选择算法对白血病、结肠癌数据集进行特征选择得出的结果，表 4-1 列出了白血病数据集上 15 个最重要的特征基因序号。表 4-2 列出了结肠癌数据集上 10 个最重要的特征基因序号。

表 4-1 白血病数据集重要基因序号

461	1745	1834	2020	2043
2242	2288	2759	3258	3320
3847	4196	4847	5039	6201

表 4-2 结肠癌数据集重要基因序号

245	249	267	365	377
493	780	1058	1771	1772

4.3.2 分类结果

实验中对于白血病和结肠癌两个数据集选取了不同数量的特征基因进行样本分类。对于训练集采用 10 折交叉验证评估训练后的分类器性能，然后对测试集进行独立测试。实验结果如表 4-3 至表 4-10 所示：

表 4-3 信噪比在白血病数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1782	97.4	76.5
891	100.0	82.4
445	100.0	85.3
200	100.0	88.2
100	100.0	79.4
50	97.4	76.5
20	97.4	79.4
10	97.4	76.5

表 4-4 t-test 在白血病数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1782	100.0	73.5
891	100.0	79.4
445	100.0	82.4
200	100.0	73.5
100	100.0	73.5
50	100.0	70.6
20	100.0	70.6
10	100.0	79.4

表 4-5 mRMR 在白血病数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1782	100.0	73.5
891	100.0	79.4
445	100.0	79.4
200	100.0	73.5
100	100.0	73.5
50	100.0	76.5
20	100.0	79.4
10	100.0	73.5

表 4-6 SVM-RFE 在白血病数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1782	100.0	73.5
891	100.0	79.4
445	100.0	79.4
200	100.0	79.4
100	100.0	79.4
50	100.0	73.5
20	100.0	82.3
10	100.0	88.2

表 4-7 信噪比在结肠癌数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1000	96.8	87.1
500	96.8	83.9
200	96.8	83.9
100	96.8	83.9
50	93.5	83.9
20	93.5	80.1
10	87.1	80.6

表 4-8 t-test 在结肠癌数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1000	96.8	87.1
500	96.8	80.6
200	96.8	83.9
100	96.8	83.9
50	96.8	77.4
20	93.5	71.0
10	93.5	64.5

表 4-9 mRMR 在结肠癌数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1000	96.8	83.9
500	96.8	83.9
200	96.8	83.9
100	96.8	87.1
50	90.3	83.9
20	93.5	77.4
10	93.5	83.9

表 4-10 SVM-RFE 在结肠癌数据集上的结果

特征数目	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
1000	96.8	87.1
500	96.8	87.1
200	96.8	83.9
100	96.8	87.1
50	93.5	83.9
20	90.3	83.9
10	90.3	83.9

在白血病数据集中，使用 t-test、mRMR、SVM-RFE 三种方法进行特征选择后，SVM 分类器在训练集的 10 折交叉验证都达到了 100%的样本识别率，证明算法选择的特征基因都是对于分类比较重要的基因。而信噪比方法在选择 100 个特征基因训练分类器才达到 100%的样本识别率。对于在测试集上的独立检测，SVM-RFE 方法有着显著的效果，在选择了 10 个特征基因的基础上进行分类就能达到 88.2%的准确率。其效果优于其它三种方法。其它三种方法在选择少量特征基因的分类效果差不多，信噪比方法在选择了 200 个特征基因后达到了分类最佳值 88.2%，而 t-test 方法是 445 个，准确率为 82.4%。SVM-RFE 方法效果显著，但运行时间远大于其它三种方法。

在结肠癌数据集中，mRMR、SVM-RFE 方法在选择了 50 个以下的特征基因得到的性能优于信噪比和 t-test 方法。四种方法在选择了 100 个特征的基础上都达到了最佳分类准确率，mRMR、SVM-RFE 为 87.1%，信噪比和 t-test 方法为 83.9%。值得一提的是在结肠癌数据中，虽然选择了 1000 或 500 个基因进行分类器训练也得到了比较好的分类结果，但是实验目标是选择最少的基因达到最佳的分类准确率。所以该数据不具有实际意义。

4.3.3 混合法和简单集成方法实验

为了选取尽可能少的特征基因且获得较高的分类准确率，实验过程中混合了 t-test 和 SVM-RFE 两种特征选择方法。先用 t-test 过滤掉一部分无关基因，再用 SVM-RFE 算法进行特征选择，将排序系数较高的 10 个基因作为特征子集。同时根据以上 4 种典型的特征选择算法得出的结果，使用一种简单的集成方法，将出现频率较高的 10 个基因作为特征子集。两种方法在两个数据集上的实验结果如表 4-11 和表 4-12 所示。

表 4-11 两种方法在白血病数据集上的结果

特征选择方法	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
混合法	100	85.3
简单集成方法	97.4	85.3

表 4-12 两种方法在结肠癌数据集上的结果

特征选择方法	样本识别率/%（10-折交叉验证）	样本识别率/%（独立测试）
混合法	90.3	83.9
简单集成方法	93.5	80.6

实验证明混合法和简单集成方法在选取较少的特征基因下能获得较好的分类准确率。因此在实际问题中，可以考虑使用这两种方法的思想进行特征选择。

4.3.4 结果的生物学分析

本节对选择的特征基因在生物学的角度作出分析。对于白血病数据集，4847 号基因(X95735, Zyxin)被证实与白血病细胞有关，它为细胞粘性蛋白编码。1834 号基因也与白血病有关，在其它论文中也被选择出来。对于结肠癌数据集，1772 号基因 (H08393, Collagen alpha 2(XI) chain)与细胞

粘性的胶原有关, 而胶原在癌细胞的新陈代谢的过程中会发生降解。249 号基因(M63391) 也被证实与肿瘤有关，肌间线蛋白是结肠癌基因中已知的枢纽癌基因之一。493 号基因、1053 号基因等在其它文章中也被选出来。

.

第5章 结论

5.1 本文研究总结

随着生物信息学的迅速发展, 传统的模式识别技术已无法满足处理高维小样本数据中大量不相关特征的要求, 提高特征选择算法的性能变得越来越重要。另一方面, 特征选择在机器学习与数据挖掘等领域的研究已取得了一些研究成果, 特征选择技术移植性地在生物信息学这一领域的研究发展将是一个有潜力的、值得深入研究的方向。与统计学、信息科学等学科的合理交叉，为微阵列数据信息的提取提供最佳路径。针对生物信息学中的海量高维小样本数据的处理, 本文总结了特征选择方法的基本研究情况及其在数据分析方面成功的应用。随着生物信息学的进一步研究与发展，产生的数据会越来越多，问题的数据也会呈现更多样化，样本数据也会随着增长，特征选择方法在生物信息学的应用也会越来越多，研究其在生物信息学中的应用，仍将会是未来的一大热点。

5.2 未来工作的展望

微阵列数据特征选择与分类是一个博大精深的研究领域，由于本人知识和时间有限，只研究了特征选择算法的冰山一角，对于算法的精髓还理解得不够深入，很难提出改进算法。下一步研究将集中在以下几个方面：

1. 集成学习最近几年发展迅猛，已经成为机器学习领域研究的热点问题。随着样本维数的增长，增强最终选择的特征子集的鲁棒性将会变得非常重要。集成学习独有的优势对处理小样本微阵列数据非常重要。混合合适的评估准则建立鲁棒的特征选择模型将会是特征选择研究的一个有趣的方向。

2. 随着深度学习在计算机视觉领域的成功，人们越来越关注数据的内部结构。如何将深度学习应用到生物信息学领域对数据进行无监督的特征学习将会是关注的重点。

3. 本文主要研究基于二分类问题，对于多分类问题的特征选择还有待深入研究。

致　　谢

本科的学习和生活马上就要结束了，我也终于完成了我的毕业论文《生物信息学中的特征选择方法总结与实现》。在这四年的本科生活中，我经历了很多学习生活中的酸甜苦辣。在此论文完成之际，谨向曾经指导、关心和帮助我的良师益友们表达我最衷心的感谢。

首先，我要感谢我的导师丰小月老师。没有她的帮助，我想我是没有办法完成这篇论文的。老师给我定的论文题目使我对科研产生了浓厚的兴趣。在论文的写作过程中，丰老师给我提出了很多建设性意义的指导意见，并解决了我对于生物信息领域问题的许多困惑。在此，再次向丰小月讲师表示衷心的感谢和诚挚的敬意。

感谢 2009 级计算机科学与技术学院四班的所有同学，是你们让我度过了快乐的本科生活，感谢你们四年时间里对我学习和生活上的关心与帮助，在此离别之际，对你们表示由衷的谢意。

最后深深感谢我的家人，他们在我的成长道路上付出了极大的心血，感谢他们多年来对我生活上无微不至的关怀、精神上的鼓励、学业上的支持。

谨以此文献给所有关心、帮助我的亲人老师和同学们！

参考文献

[1] Dash M, Liu H. Feature selection for classification[J].SIntelligent data analysis, 1997, 1(1-4): 131-156.

[2] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics[J]. Bioinformatics, 2007, 23(19): 2507-2517.

[3] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of bioinformatics and computational biology, 2005, 3(02): 185-205.

[4] 邵进智. 基于属性间相关性分析的属性选择方法研究[D]. 北京交通大学. 2009.

[5] Hauskrecht M, Pelikan R, Valko M, et al. Feature selection and dimensionality reduction in genomics and proteomics[M]. Fundamentals of Data Mining in Genomics and Proteomics. Springer US, 2007: 149-172.

[6] 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.

[7] 梁爽. 基于 SVM-RFE 算法的癌症基因表达数据分析[D]. 吉林大学. 2012.

[8] 游伟. 基于支持向量机的基因选择算法研究[D]. 湖南大学. 2010.

[9] 梁艳春，张琛. 生物信息学中的数据挖掘方法及应用[M]. 北京：科学出版社，2011，125-179.

[10]于化龙，顾国昌. 基于 DNA 微阵列数据的癌症分类问题研究进展[J]. 计算机科学，2010，16-21.

[11]Yang P, Hwa Yang Y, B Zhou B, et al. A review of ensemble methods in bioinformatics[J]. Current Bioinformatics, 2010, 5(4): 296-308.

[12]Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization

and genetic algorithm[J]. Soft Computing, 2008, 12(11): 1039-1048.

[13]Liao C, Li S, Luo Z. Gene selection using wilcoxon rank sum test and support vector machine for cancer classification[M].Computational Intelligence and Security. Springer Berlin Heidelberg, 2007: 57-66.

[14]Li S, Liao C, Kwok J T. Gene feature extraction using T-test statistics and kernel partial least squares[C].Neural Information Processing. Springer Berlin Heidelberg, 2006: 11-20.