吉林大学

本科生毕业论文（设计）

中文题目 基于随机重启的递增式特征选择算法研究

英文题目 Incremental Feature Selection with Random Re-start

学生姓名 叶玉婷 班级 211214 学号 21121423

学　　院 计算机科学与技术学院

专　　业 物联网工程

指导教师 周丰丰 职称 教授

吉林大学学士学位论文（设计）承诺书

本人郑重承诺：所呈交的学士学位毕业论文（设计），是本人在指导教师的指导下，独立进行实验、设计、调研等工作基础上取得的成果。除文中已经注明引用的内容外，本论文（设计）不包含任何其他个人或集体已经发表或撰写的作品成果。对本人实验或设计中做出重要贡献的个人或集体，均已在文中以明确的方式注明。本人完全意识到本承诺书的法律结果由本人承担。

学士学位论文（设计）作者签名：

2015 年 5 月 20 日

基于随机重启的递增式特征选择算法研究

摘要

大数据时代的到来催生数据建模和分析的研究，现有一个重要的研究热点——基于大数据的癌症标记物的健康诊断，具有广阔的应用前景，对人类健康具有重要的意义。在超高维的特征集合中，穷举搜索最优的特征子集的搜索计算时间开销会随特征集合的维度呈指数爆炸性增长，此问题已被证明为 NP 难问题。所以运用传统特征选择方法在高维特征集合中选择最优特征子集在计算上是不可行的。为分析海量数据，多种特征选择算法应运而生，特征选算法的研究呈现出百家争鸣的姿态。特征选择，即在复杂冗余的高纬数据中，剔除不相关的冗余数据，挑选出特征个数尽量少，类别相关性强的最优特征子集，是一个降维的过程。本文提出基于随机重启的递增式特征选择算法（RIFS, Incremental Feature Selection with Random Re-start），在递增式特征选择算法上提出随机重启的思想，通过随机产生的 m 个数据，作为特征递增排序后的相应起始位置，从各起始位置往下依次找 k 个子特征作为一个特征子集，这样的特征子集一共有 m 个。最后利用 K 折交叉验证和封装式（wrapper）特征选择，挑选出最优的特征子集。本文通过对 16 个大数据集的测试，发现基于随机重启的递增式特征选择算法具有很好的表现且选出的特征子集的特征个数相对较少。结果表明，从特征选择效果和特征子集特征数目综合来看，RIFS 算法能够正确地选择特征，高效地去除冗余特征。

关键字：特征选择，随机重启，K 折交叉验证，wrapper 方法，filter 方法

Incremental Feature Selection with Random Re-start

Author: Yuting Ye

Tutor: Fengfeng Zhou

Abstract

The advent of Big Data Era had led to various researches about mathematical modeling and analysis. The big data based biomedical diagnosis of cancer biomarkers has broad applications and important implications for human health. And it is becoming a major research focus. Selecting the optimal subset of features from a high-dimensional dataset using traditional feature selection algorithms is computationally infeasible, because the health big data has an exponentially increased dimension. Feature selection has proven to be an NP-hard problem, which means it's impossible to find a global optimal solution within a reasonable time. In order to analyze huge amounts of big data, a variety of feature selection algorithms emerged. Feature selection is a dimension reduction process aiming at weeding out the irrelevant redundant dimensions from a complicated and redundant high dimensional dataset. This work describes a feature selection algorithm RIFS based on IFS (Incremental Feature Selection). It presents the idea of random re-starts. In this work, multiple randomly generated numbers will be used as the starting positions of the ascending sorted feature subsets. From each of the starting positions, followed by the selection of a fixed number of features, as a feature subset. Each of such feature subsets will be validated by the 3-fold cross validation using 4 classifiers. In this work, RIFS is applied on 16 biomedical datasets. According to the experimental results, and this algorithm outperforms most of

the existing algorithms. The features selected by RIFS also tend to have high biomedical correlation with disease progression functions. RIFS also tends to select much smaller feature subsets than the existing algorithms, while achieving similar classification performances. We believe that RIFS represents a novel feature selection strategy that may facilitate the computational disease diagnosis for the clinicians.

Keywords: Feature selection, random re-start, k-fold cross validation, filter algorithm, wrapper algorithm.

目　　录

第 1 章 绪论

1.1 研究意义和目的

癌症是目前威胁人类健康最为危险的杀手之一，而在各类癌症中，肺癌引起了社会广泛的关注。引发肺癌的因素诸多，如环境因素—雾霾等，人类活动因素—吸烟，遗传因素等，使得诱发肺癌的机率持续走高，更有研究表明，我国或成为罹患肺癌的第一大国。

现今对癌症的预防和治疗并不乐观，不少民众发现罹患癌症之时或以为时过晚，因此癌症也成为了最为可怕的病症之一。然而癌症肿瘤与人类基因密不可分。癌症是人类生活内因（基因）和外因（环境，饮食，活动）共同作用的结果，通过基因检测，癌症标志物检测等方法，能够预先知道身体中可能存在的疾病的潜在隐患，从而能够得到及时的积极治疗。

因此，寻找一个通过基因检测的方法来预测人体罹患癌症的机率，对癌症肿瘤的检测治疗具有深远意义。本课题研究采用基于随机重启的递增式特征选择算法研究寻找与肺癌相关性高基因组，从而达到高效检测肺癌的目的。

1.2 国内外现状

目前由于环境污染等因素，越来越多的疾病正向人们的生活靠近。而癌症成为威胁人类健康最为可怕的疾病之一，如何防治癌症成为一个研究热门话题。癌症致病基因的发现与预测对认识肺癌发生的机理及其诊断防治有着重要意义。因此如何从众多基因中选择出癌症相关的基因显得尤为重要。在这方面国内外学者至今做了许多研究。

在众多癌症基因选择的算法中有如下具有代表性的几种算法，启发

式宽度优先搜索[1]，基于 RD-SVM 的选择算法，基于二元网络异步重启随机游走算法[2]，基于递增式特征选择算法[3]等等。这些算法中基于递增式特征选择算法在癌症标志物检测中运用的相对较多。但是对基于随机重启的特征选择算法在癌症标志物的检测中还未出现或可能有此想法但并未深入研究。所以该技术是一个较为新颖的技术。研究此技术对癌症防治，医疗健康具有重大意义。

1.3 技术简介

本课题使用R语言作为开发语言，R版本为R version 3.2.2(Fire safety)，集成开发环境为 Rstudio。R 语言是一种自由软件编程语言与操作环境，通常用于统计分析，绘图以及数据挖掘。

R 内置多种数字及统计学分析功能。R 的功能可通过安装包（Packages）增强。R 是 S 语言的一个分支，可认为是 S 语言的一种实现。相对于其他数学或统计学专用的编程语言，R 具有更强大的面向对象功能。R 的另一强项是优秀的统计制图功能。R 不仅用于统计分析和开发统计相关的软件，也有人用作矩阵计算。在性能上，其表现也相当完美，分析速度与专用于矩阵计算的自由软件 GNU Octave，商业软件 MATLAB 不相上下。

R 语言具有良好的发展和应用前景，生物信息学社区常使用 R 进行生物学数据分析。R 也被选作为 Bioconductor 计划中基因图谱分析所使用的工具。

1.4 创新性与主要贡献

RIFS 算法相对与传统的特征选择算法能提取出性能更优的特征。在实际运用中适用领域广泛，在各个有关大数据的领域中发挥着不容小觑的作

用，例如，在人类健康方面该算法能用于提取癌症基因同时配合其它算法实现对人体健康情况的预测。算法的优良表现来源于加入随机重启的基本思想，这一创新性的举措让算法在搜索特征子集时搜索范围更广使获得性能更优的特征子集的可能性大大提高。RIFS 算法的创新性与贡献主要分为以下三条。

第一条， RFIS 算法保留了 IFS 的优越性。

第二条， RIFS 算法在 IFS 算法的基础上，扩大了候选特征子集的提取范围，使得算法的精确性，准确度更高。

第三条， 实验结果表明，RIFS 算法在未来癌症标志物的预测方面具有广阔的应用前景。

第 2 章 特征选择算法

2.1 特征选择

特征选择是从一组特征中挑选出一些最有效的特征以降低特征空间维数的过程[4]。从形式上来说，特征选择就是，在一个特征数为 N 的大数据集中，选出一个特征数目为 M 的最小特征子集。该特征子集是原特征集剔除了不相关的，冗余特征的结果，从而提高精度，减少运行时间。特征选择的结果直接影响着分类器的精度和性能。经典特征选择方法是从高纬特征集合中选取一些特征，这个方法包括了特征选择和特征提取两方面。其中广义上的特征提取是寻找处于高维度的特征样本和低维空间特征集合的一种映射关系，从而达到降维目的；特征选择还可以解释为从原始特征集合中剔除冗余的或者不相关的子特征。这两者常配合使用，首先通过特征提取的方法将高维特征映射成低维特征，完成该映射后再通过特征选择去除冗余的不相关特征，进一步达到降维的目的。

2.2 特征选择过程

特征选择主要分为子集产生，子集评估，停止条件，子集验证四个过程。其中子集产生是指根据相应的特征算法产生待筛选的特征子集。子集评估是指按照某个评价函数准则评估该特征子集。停止条件用于决定特征选择过程停止的时间。子集验证用于验证最终所选子集的有效性。Dash 等人[5]给出了如下图 2-2 所示的特征选择基本框架。

特征选择的步骤

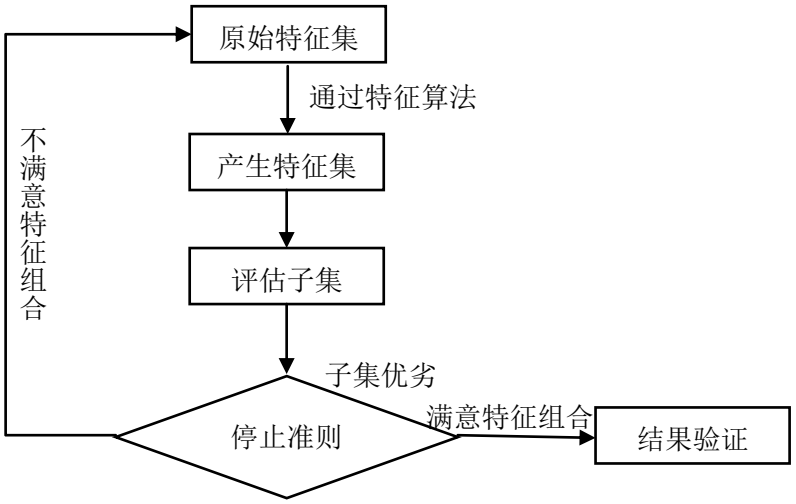


图 2-2 特征选择过程

2.3 特征选择算法

以特征选择过程中候选特征子集的评估准则以及后续所融合的学习、分类算法等为分类标准，特征选择算法可以大致分为过滤式(Filter) 、封装式(Wrapper)和嵌入式(Embedded)三种。

2.3.1 嵌入式特征选择

嵌入式可以说是一种新兴的特征选择模型，它结合学习器来评价特征子集 ，且具有封装式特征选择模型的精度和过滤式特征选择模式的效率。

2.3.2 过滤式特征选择

过滤式(filter)特征选择与特定的预测模型无关，它通过统计方法为数据集中每一个特征分配一个分值，以分值为评判标准，如按照分值将特征

排序，然后决定去除还是保留某特征。通常认为，分值较大的特征以及特征子集在后续分类算法中能够得到较理想的准确率。由于此算法不依赖于特定的预测模型如分类器，具有良好的通用性，算法的复杂性低。因此计算对象为大规模数据集时，此算法能快速去除大量冗余的不相关特征，所以非常合适作为特征的预筛选器。因为算法的评价标准独立于特定的学习算法，所选的特征子集在分类准确率方面通常比 Wrapper 方法低。

2.3.3 封装式特征选择

封装式特征选择通常利用学习算法的性能评价特征子集优劣。Wrapper 方法首先需要训练一个分类器，根据分类器的性能对一个待评价的特征子集进行评价。在 Wrapper 方法中用来评价特征的学习算法有多种，常见的有决策树、贝叶斯分类器、近邻法、神经网络和支持向量机等。相比于 Filter 方法，Wrapper 方法通常能找到分类性能更好的特征子集。

2.4 分类算法

分类是数据挖掘中的一个重要课题。分类的目的是学会一个分类函数或分类模型，此函数定义了数据集合与给定类别集合的映射关系。分类，简单来说，就是根据文本的特征或属性，划分到已有的类别中。分类可描述如下：输入数据，即训练集（Training Set），在本文中是多个子特征组成的特征向量，训练集中每个子特征均有与之对应的特定类标签(Class Label)。分析输入数据，通过训练集表现出来的数据特征，为每个类找出一种准确的模型，预测出该训练集对应的分类。在本课题中，应用了决策树分类算法(Decision Tree)，朴素贝叶斯算法(Naive Bayes),支持向量机算法(SVM，Support Vector Machine),K-临近算法(KNN,K-Nearest Neighbors)。

2.4.1 决策树分类算法(Decision Tree)

决策树算法是数据挖掘中的重要技术之一，决策树算法，顾名思义，是一种依托于树形结构建立起来的预测模型，其中树形结构的建立依托于策略的抉择。决策树一般分为三个部分，决策节点，分支和叶子节点。以根节点为首的决策节点代表一个问题或者条件，每个分支代表其父节点所提出条件的属性值，每个叶子节点代表沿根节点开始，遇到不同的决策节点进行分支而得到的分类结果。在整个树形结构中，叶子节点代表一种分类结果，非叶节点均代表问题或条件。

如图 2-4-1 是一个简单的决策实例，以某名牌大学甄选硕士研究生夏令营营员为例，非叶子节点代表评判学生的条件，叶子节点代表学生能否参加夏令营。

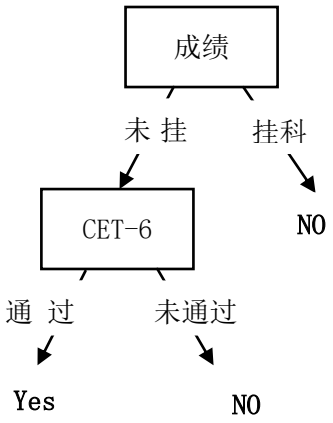


图 2-4-1 决策树实例

其中成绩未挂科并且英语过了六级的学生具有参加夏令营的资格，而挂科或者六级没过的学生均无资格参加该校的硕士研究生夏令营。

2.4.2 朴素贝叶斯算法(Naive Bayes)

贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，

故统称为贝叶斯分类，朴素贝叶斯分类(N-Bayes)是贝叶斯分类中最简单的一种。贝叶斯定理作为贝叶斯分类的最基本定理，解决了生活中很常见的问题：已知在事件 B 发生的条件下 A 发生的概率，如何求得事件交换后的概率，即在事件 A 发生的条件下，B 发生的概率呢？用下公式描述贝叶斯定理：

$$\mathrm{P(B|A)} = \frac{P(A|B)P(B)}{P(A)}$$

朴素贝叶斯的思想基础是这样的：对于给出的待分类项和已知的类别集合，求解在此待分类项出现的条件下，各个类别出现的概率，概率最大的类别被挑选出来，作为该项的最终分类。

（1）设 $x=\{a_1,a_2,\ldots,a_m\}$为一个待分类项,其中每个 a 为 x 的一个特征属性。

（2）现有类别集合 $\mathrm{C}=\{y_1,y_2,\ldots,y_n\}$，计算 $\mathrm{P}(y_1|x),\mathrm{P}(y_2|x),\ldots,P(y_n|x)$。

（3）根据$\max\{P(y_1|x),P(y_2|x),\ldots,P(y_n|x)\}$找出最大概率类别，待分类项 x 根据最大概率分为$y_k$类。

朴素贝叶斯分类方法简单，速度快，准确率高，能运用到大型数据库中。

2.4.3 支持向量机算法(SVM)

支持向量机（Support Vector Machine，常简称为 SVM）属于一般化线性分类器，广泛应用于统计分类,回归分析等,是一种监督式学习的方法。支持向量机不仅能够最小化经验误差，也能最大化几何边缘区，被称为最大边缘区分类器。定义一个良好的分类边界是训练数据集分类的必要条件。支持向量机通过建构一个或多个高维的超平面作为分类边界来划分数据点。判断分类边界的好坏需要知道分类边界与最近数据点的距离。该距离越远，分类器的泛化误差越低。

在支持向量机中，分类边界与离其最近数据点的距离称为间隔

（margin）。SVM 将向量映射到一个高维空间，并建立此空间里间隔最大的分隔超平面。然后在分类数据点的超平面两边分别建立互相平行的且距离最大化的超平面。SVM 的目标为找出最优分类边界，该最优分类边界即为间隔最大的分隔超平面。SVM 它本质上即是一个分类方法，用 w^T+b 定义分类函数，为寻最大间隔，引出 1/2||w||^2，继而引入拉格朗日因子，化为对拉格朗日乘子 a 的求解（求解过程中会涉及到一系列最优化或凸二次规划等问题），本文不再一一赘述。支持向量机实质上是非线性分类器它的学习策略是间隔最大化。

2.4.4 K-邻近算法(KNN,K-Nearest Neighbors)

K 近邻分类算法（K-nearest neighbor，KNN）是一种经典的分类器算法。想对一个未知的数据做分类，最简单的方法是遍历所有的训练数据，寻找与新数据属性相同的训练数据，若找到匹配的训练数据，则将该数据的分类作为新数据的分类。这种分类方式有一个缺点，那就是很可能找不到一个完全匹配的训练数据导致无法分类。KNN 算法则是在“物以类聚，人以群分”的思想上发展起来的基于距离的分类器。寻找完全匹配的训练数据不可行，那就找出距离待分类数据最近的 K 个邻居进行分析。从训练集中找到和待分类数据最接近的 k 条记录，然后以这 K 条记录的已知分类为依托来决定新数据的类别。KNN 分类算法一般可以分为三个步骤。首先计算待分类数据集与训练集中每个数据的距离，距离衡量一般包括欧式距离，马氏距离，夹角余弦等。接着寻找距离最近的 k 个训练数据作为邻居。最后根据这 K 个邻居的属性类别，为未知数据对象分类。

K 近邻分类算法优点有简单，无需估计参数，易于理解，无需训练。

2.5 IFS 算法

Incremental Feature Selection 算法简称 IFS，即递增式特征选算法，常用来粗略考查建立有效模型所需要的变量数目。第一步根据特征集S建立M个特征子集，即：$S=V_1$, $S_2=V_1,V_2$,…… , $S_m=V_1,V_2,\cdots\cdots,V_m$ 当中 V_i 是第i个加入到特征子集 S_i 中的特征；然后结合适合的数据挖掘方法对前面得到的特征子集分别建模。也就是从特征数分别为 1，2，3……m-1,m 的子集 $S_1,S_2,\ldots.S_m$ 做交叉验证，计算各自交叉验证预报正确率；从结果中挑选出正确率最高的特征子集 S_n 作为最终的特征子集建立模型。最后把各特征子集的交叉验证预报正确率对特征数作图即可得到所谓的 IFS 曲线。

第 3 章 RIFS 算法

3.1 二元分类以及性能指标

分类任务的目标是寻找一个函数，把观测值匹配到相关的类或者标签上。学习算法必须用成对的特征向量和对应的标签来估计匹配函数的参数，从而实现更好的分类效果。二元分类（binary classification）中，分类算法需要为一个实例配置两个类别。二元分类案例包括，预测患者是否患有某种疾病，邮件是否为垃圾邮件等。我们探讨二元分类问题时需要两个样本集，现有样本中特征个数分别为 n 和 m 的样本集 P 和样本集 N，其中 P 为正类集(Positive Set),N 为负类集(Negative Set)。现在有一个特征集 X，且 X 中每一个特征都包含在 P 或 N 中，二元分类要做的事情就是，根据一定的分类算法，判别特征集 X 是属于正类还是负类。

对一个二分问题来说，会出现四种情况。如果一个实例是正类并且也被预测成正类，即为真正类（True positive）TP,如果实例是负类被预测成正类，称之为假正类（False positive）FP。相应地，如果实例是负类被预测成负类，称之为真负类（True negative）TN,正类被预测成负类则为假负类（false negative）FN。

通常用三个指标来判断一个二分模型性能的优良，分别为 Sensitivity (Sn), specificity (Sp) and Accuracy (Acc)，其中 Sn=TP/(TP+FN)，Sp=TN/(TN+FP)，Acc=(TP+TN)/(TP+FN+TN+FP)。在本课题中，选择具有更高 ACC 的特征子集并且特征数少的特征子集作为最优选择。

3.2 交叉验证

交叉验证技术常用于评价一个统计分析的结果能否推广到一个独立

的数据集。通常用于评估一个预测模型在实际应用中的准确度。在统计学上常用交叉验证技术将数据样本切割成较小子集。这使得我们可以先在一个子集上做分析，然后利用其他子集对此分析做后续的确认以及验证。一个交叉验证需要将样本数据集划分成两个互补的子集，其中一个子集用于训练（分类器或模型），将其称为训练集（training set）；而另一个子集称为测试集（testing set），它用于验证（分类器或模型的）分析的有效性。利用测试集来测试分类器或模型，最终的测试结果作为性能指标。交叉验证的研究期望是得到较高的预测精确度和尽量低的预测误差。为了提高交叉验证结果的准确度，一般会多次划分将已知样本数据集，得到不同的互补子集，进而进行多次交叉验证。求取多次验证的平均值并作为交叉验证的结果。

K 次交叉验证，顾名思义，将初始样本数据集分割成 K 个子样本数据集，取其中一个子样本作为测试集，其余 K-1 个子样本用作训练集，交叉验证重复 K 次，每次变换不同的子样本作为测试集，以保证每个样本验证一次。最终取 K 次交叉验证结果的平均值作为验证结果。以 10 折交叉验证为例，现有样本数据集 A，将 A 分成割 10 个子数据集得到 $\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}\}$。先以子集 A_1 作为测试集，以 $\{A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}\}$作为训练集，做交叉验证得到一个验证结果 M_1；再以 A_2 作为测试集，剩余$\{A_1, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}\}$作为训练集，做交叉验证得到验证结果 M_2；同理，下来分别以 A_3 A_4 A_5…A_{10} 作为测试集，其余剩余部分分别作为相应的训练集得出验证结果 M_3 M_4 M_5 …M_{10}，求平均后 M= $\frac{M_1+M_2+\cdots M_{10}}{10}$是最终的验证结果。

3.3 应用生物医学数据集介绍

本课题中应用了 16 个二元分类数据集作为分类评价的基础数据集。其中数据集 *Colon*[6]提取自 R package ColonCA，数据集 *Leukaemia*[12]提取自 Bioconductor package golubEsets。另外 6 个常用的数据集 *DLBCL*[7]，*Prostate*[8]，*ALL*[9]，*CNS*[10]，*Lymphoma*[11]下载自 Broad Institute Genome Data Analysis Center，网页链接为 http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. 其中 *ALL* 进一步分为数据集 *ALL1*，*ALL2*，*ALL3*,*ALL4*，还有 6 个数据集 *Myeloma* (Accession: GDS531) [13], *Gastric* (Accession: GSE37023) [14], *Gastric1*/*Gastric2* (Accession: GSE29272) [15], *T1D* (Accession: GSE35725) [16] and *Stroke* (Accession: GSE22255)[17],下载自 NCBI Gene Expression Omnibus (GEO)数据库。下表 3-3 反映了各个数据集样本数和特征数。

ID	Dataset	Samples	Features
1	DLBCL	77	7129
2	Pros (Prostate)	102	12625
3	Colon	62	2000
4	Mye (Myeloma)	173	12625
5	ALL1	128	12625
6	ALL2	100	12625
7	ALL3	125	12625
8	ALL4	93	12625
9	CNS	60	7129
10	Lym (Lymphoma)	45	4026
11	Leuk (Leukaemia)	72	7129
12	Gas (Gastric)	65	22645
13	Gas1 (Gastric1)	144	22283
14	Gas2 (Gastric2)	124	22283

续表

ID	Dataset	Samples	Features
15	T1D	101	54675
16	Stroke	40	54675

表 3-3 测试数据集

3.4 基于随机重启的特征选择算法设计

3.4.1 问题描述

现有 16 个关于癌症标志物的高纬样本数据集，如何在这 16 个数据集中提取出各自和癌症相关性最优的特征子集，是本科题需要解决的问题，而基于随机重启的特征选择算法则为本课题提出的解决该问题的算法。

关于各个高纬样本数据集，我们不难得到特征集合中每个子特征癌症相关性大小。可以想象，相关性越大的子特征越容易被挑选出来，然而最优特征子集一定是由相关性均靠前的子特征组合而成吗？排名靠后的子特征对最优特征子集的提取有贡献吗？因此我们做出以下猜想，排名靠后的特征组合也可能具有良好的表现，提出随机重启的思路，根据随机生成的 m 个数据，在不同的排名位置寻找最优子集。

3.4.2 算法设计思路

t-检验可以通过分布理论推论差异发生的概率，从而得出两个平均数的差异化的大小。本算法中，首先对 ALL 等 16 个输入数据集做 t-检验，得出 P-value 代表差异是否显著。根据 P-value 值对各个数据集中子特征排序，大致认为，排序靠前的特征与癌症的相关性更大。基因组合的多样性及组合后表现与单个变现差异很大，所以认定与癌症相关性高的基因组合

是排序靠前的基因组合过于偏面。随机生成 M 个随机数字，在排序后的数据集中，分别以这 M 个随机数字为起始位置，依次往后寻找 K 个特征组成 M 个特征子集。对这 M 个特征子集，根据已有的类标(Class Label)，分别采用 D-tree，SVM，NBayes，KNN 四个分类算法进行分类，期间运用 3 折交叉验证估算分类的准确度，最终 ACC 最高的特征子集将被挑选出来。下图 3-4-2 为 RIFS 算法大致流程。

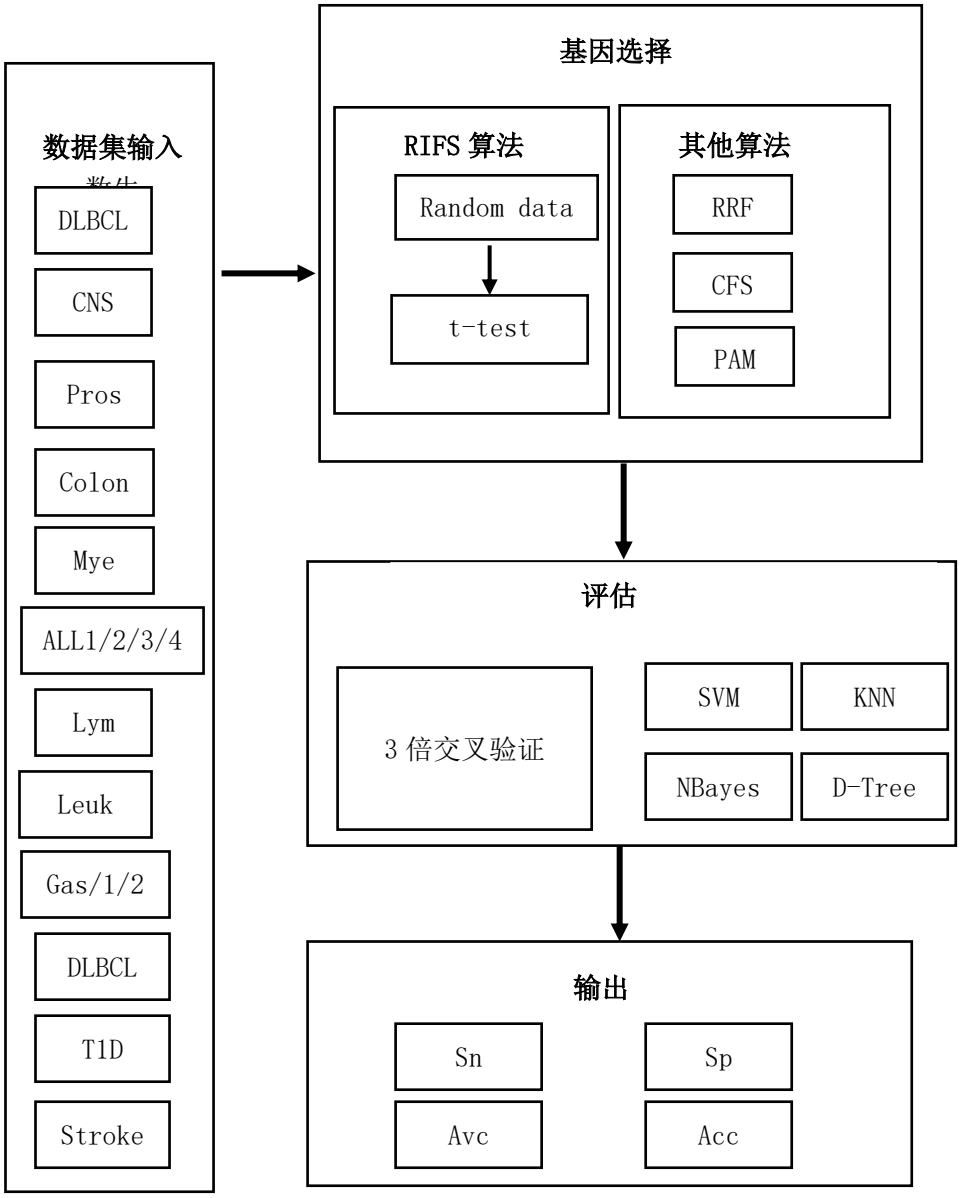


图 3-4-2 RIFS 算法流程

3.4.3 RIFS 算法的优化

提高一个算法的准确度和可行性，需要不断地优化，针对以上算法的设计思路，我们对 RIFS 算法提出了三点优化。

保存排名靠前的子特征，考虑其在特征提取中的贡献。在进行各个子特征相关性排序后，若直接根据随机生成的数据作为提取特征子集的起始位置，这样容易忽略排名靠前的子特征，因为随机生成的数据中很有可能不包含排名靠前的位置。因此，在算法过程中，将排序靠前的特征提取出来，计算其优越性，与之后的计算进行融合比较，最终挑选出最优特征子集。

1.前向搜索，扩大特征提取的搜索范围。以下图 3-4-3(1)所示 IFS 曲线为例，曲线呈递减形势，序号为 9552 的特征其计算出的 ACC 均比其后的特征组合的 ACC 要高，因此猜想，融入前一个或几个子特征后，可能会得出更好的计算结果。此优化即前向扩大搜索范围，寻求更优的特征子集。

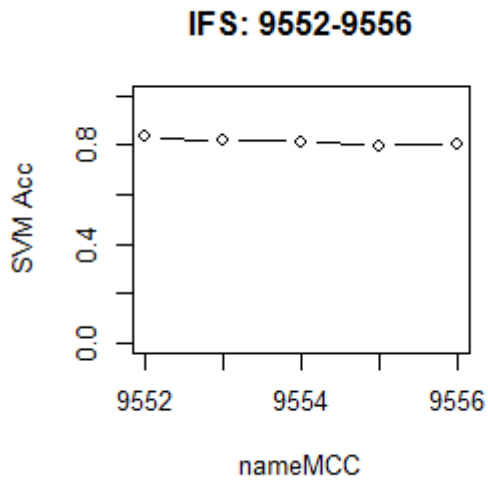


图 3-4-3(1) RIFS 图例(1)

2.后向搜索，扩大特征提取的搜索范围。以下图 3-4-3(2)所示 IFS 曲线为例，曲线的最高点出现在末尾，序号为 11-15 的子特征组成的特征子集具有最高的 ACC,因此猜想，加入序号为 16 的子特征后，计算出的 ACC 能否提高呢？继续加入序号为 17 的呢？此优化即后向扩大搜索范围，寻求更优的特征子集。

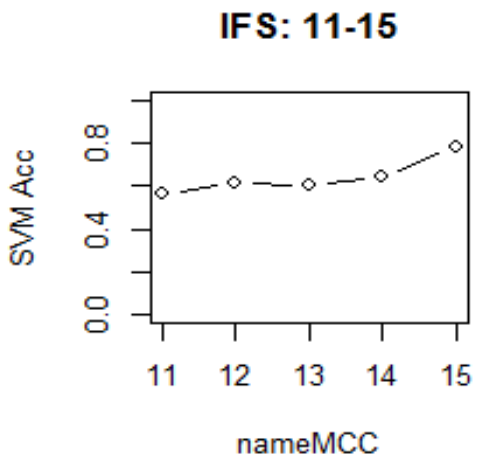


图 3-4-3(2) RIFS 图例(2)

3.5 RIFS 算法框架

本文针对求解多个生物健康大数据中关于癌症标志物的最优特征子集提出基于随机重启的递增式特征选择算法，简称 RIFS 算法。

图 3-5 给出 RIFS 算法求解最优特征子集的过程。RIFS 算法首先初始化数据集和与其对应的分类集合(ClassLable)（1 行）。然后使用 t-检验训练各数据集，训练结果保存在 dataFtest 中（2 行），dataFtest 为 n 行 2 列矩阵，n 指数据集中特征个数。矩阵第二列保存的是 t-test 后的 P-value 值（3 行），代表数据集中各个子特征的癌症相关性大小。以 PValue 的值为指标对矩阵 dataFtest 排序得到新的矩阵 egRank（4 行）。随机生成一组数 randomData(5 行)其中 numRd1 为随机生成的随机数的个数，egNumSmall 为 Dataset 中特征个数， Numnamecc1` 为候选特征子集中的特征个数，egNumSmall-Numnamemcc1 是为了防止生成随机数越界。主循环从第 6 行到第 12 行被执行，直到遍历 randomdata 中的各个数据后停止，在每次迭代中，算法使用函数 getAcc()(3.6 节)计算各个候选特征子集的性能指标保

存在 temp 里面然后若比之前保存的性能指标更优那么就更新 Maxresult 中的值替换为当前最优解。算法最后返回最优解 Maxresult（13 行）。

RIFS 算法过程

```
1、 initialize the dataset egMatrix and the ClassLable
2、 egClass;dataFtest <- apply(t.test(egMatrix, egClass));
3、 P-value <- dataFtest[ ,2];
4、 egRank <- rank(dataFtest[,2]);
5、 randomData <- round(runif(numRd1,1,egNumSmall-Numnamemcc1));
6、 for (h in randomData){
7、       temp <- getAcc(h,Numnamemcc1);
8、       if(temp[1]>Maxresult[1]){
9、            Maxresult[1]<-temp[1];
10、      Maxresult[2]<-temp[2];
11、      Maxresult[3]<-temp[3];
12、}}
13、Return Maxresult;
```

图 3-5 RIFS 算法过程

3.6 各特征子集性能指标计算 getAcc()

本节将分析计算最优性能指标的过程。在整个计算过程中需要采用 D-tree，SVM，NBayes，KNN 四个分类算法进行分类，期间运用 3 折交叉验证估算分类的准确度。具有最高 ACC 值的特征集合将被挑选出来。

图 3-6 给出 RIFS 算法中 getAcc()函数计算的候选特征子集性能指标的详细过程。首先初始化 MaxAcc，MaxPoint，NumberMaxAcc 为 0，分别保存候选特子集的最高 ACC，子集起始位置，最终挑选出来的最优子集的子

集个数（1 行）。第 2 行到 14 行 for 循环为生成 1、2…egTopK 个候选子集，并计算每个候选子集各自的性能指标。第 3 行为在原始数据集 egMatrix 上，以 egStart 为起始位置,往下寻找 i 个特征组成新的候选特征子集 tResultMatrix。第 4 行到 7 行 for 循环为对候选子集做三倍交叉验证。第 5 行表明通过 SVM,KNN,D-Tree,NBayes 四个分类算法对候选子集 tResultMatrix 分类。计算 se，sp，Acc 分析各个特征子集的性能，se，sp，Acc 计算公式可参见 3.1 节，最终挑选出具有最大 ACC 的特征子集左右待选最优特征子集。第 15 行返回待选最优特征子集的最大 ACC——MaxAcc，子集起始位置 MaxPoint，子集个数 NumberMaxAcc。

---

getAcc() 函数算法步骤

```
1、 MaxAcc<-0; MaxPoint<-0;NumberMaxAcc<-0;
2、 for( i in 1:egTopK ){        // egTopK 为候选特征子集个数，egStart 起始位置
3、 tResultMatrix <- egMatrix ( (egRank>=egStart) & (egRank<=(egStart+i-1)) );
4、     for(i in 1:3){
5、         Predy<-clssifier(tResultMatrix,”SVM”,”KNN”,”D-Tree”,”NBayes”
6、         Cv(i)<-crossvalidation（i）
7、         num_tp[i], num_fn[i], num_fp[i], num_tn[i]<-Evaluate{ Cv(i) & predy}
8、     se=sum(num_tp)/sum(tResultMatrix ==1)
9、     sp=sum(num_tn)/sum(tResultMatrix ==0)
10、     Acc=sum(num_tp+num_tn)/length(tResultMatrix)
11、     if(Acc >MaxAcc){
12、         MaxAcc<- Acc
13、         MaxPoint<-egStart;
14、         NumberMaxAcc<-i}}
15、     return (MaxAcc,MaxPoint,NumbefMaxAcc)
```

图 3-6 getAcc() 函数算法过程

第 4 章 结果分析与讨论

4.1 RIFS 算法性能

利用基于随机重启的递增式特征选择算法(RIFS)，对 DLBCL，Prostate，CNS，等 16 个数据集进行计算，经优化后，数据集 DLBCL 计算出的最大 ACC，下文中用 mAcc 表示，值为 0.974，特征子集中特征个数为 4。数据集 Prostate 计算出 mAcc 的值为 0.953，特征子集中特征个数为 5。数据集 Colon 计算出 mAcc 的值为 0.919，特征子集中特征个数为 4。因有 16 个数据集，在此不一一叙述，16 个数据集提取出的最优特征子集所计算出的 mAcc 以下图形势给出。

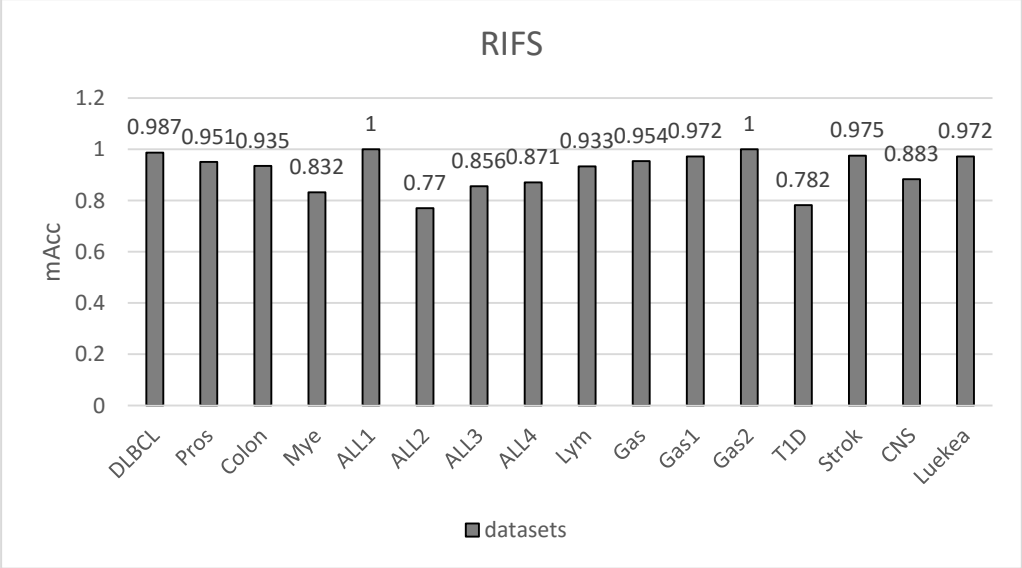


图 4-1 各数据集 mAcc

4.2 算法性能比较

4.2.1 算法结果

为对比阐述基于随机重启的递增式特征选择算法的性能，本文还引入了其他三种现有的特征选择算法，分别是 PAM[18]，RRF[19]，CFS[20]，算法定义本文将不再一一赘述。为清晰的表明算法之间性能的比较，各个算法在不同数据集上的最优性能的数据将以表格的形式给出，见表 4-2-1(1)和表 4-2-1(2)阐明各个算法在不同数据集上提取的最优特征集合中子集的个数。

	DLB	Pros	Colon	Mye	ALL1	ALL2	ALL3	ALL4
RIFS	0.987	0.951	0.935	0.832	1	0.77	0.856	0.871
CFS	0.987	0.949	0.919	0.899	1.000	0.837	0.838	0.948
PAM	0.966	0.920	0.916	0.853	1.000	0.651	0.805	0.915
RRF	0.943	0.864	0.886	0.794	0.975	0.666	0.808	0.885
	Lym	Luek	Gas	Gas1	Gas2	T1D	Strok	CNS
RIFS	0.933	0.972	0.969	0.972	1	0.782	0.975	0.883
CFS	1.000	1.000	0.975	0.972	0.988	0.905	1.000	0.843
PAM	0.995	0.986	0.941	0.949	0.977	0.716	0.793	0.771
RRF	0.943	0.932	0.892	0.957	0.988	0.698	0.958	0.643

表 4-2-1(1) 各算法性能指标(1)

	DLBC	Pros	Colo	Mye	ALL1	ALL2	ALL3	ALL4	CNS
RIFS	9	5	6	1	8	3	1	6	5
CFS	141	80	53	107	103	56	43	105	39
PAM	52	2	14	34	1	1	2	30	20
RRF	12	14	16	39	10	31	34	16	29
	Lym	Leuk	Gas	Gas1	Gas2	T1D	Strok		
RIFS	1	9	2	9	12	5	4		
CFS	68	39	115	300	237	166	138		

PAM	109	20	44	29	500	48	3
RRF	9	29	7	16	9	28	14

表 4-2-1(2) 各算法性能指标(2)

根据表二中数据，RIFS 算法在在不同数据集上的平均最优性能 mAcc 比 CFS 算法低 2.3%，比 PAM 算法和 RRF 算法分别高出 3.5%和 5.9%。尽管根据表 2，RFIS 算法的平均性能相对 CFS 算法低，但 RIFS 算法提取出的特征子集中特征总共个数仅为 CFS 的 4.8/100，明显少于 CFS 算法。以两种算法在 ALL1 数据集上的性能比较为例，RIFS 和 CFS 在 ALL1 上的最优性能指标 mAcc 值均为 1，而 CFS 算法提取的特征子集中的特征个数为 103，而 RIFS 算法的特征个数仅为 8。由此可以看出，RIFS 算法不仅能够提取出性能较好的特征子集，且特征子集明显具有特征个数低的优点。

4.2.2 各算法 mAcc 比较

本课题中判别特征选择算法的一项重要指标为特征子集的最大 Acc, 即 mAcc。下表详细给出各个特征选择算法在 16 个数据集上最大 Acc 的比较。以 RIFS 算法和 CFS 算法的比较为例，RIFS 算法在数据集 Prostate，Colon，ALL3，Gas2，CNS 计算出的最大 Acc 值分别为 0.951，0.935，0.856，1，0.883，大于 CFS 算法分别计算出的 0.949，0.919，0.838，0.975，0.843。RIFS 算法在数据集 DLBCL，ALL1，Gas1 上计算出的最大 Acc 值与 CFS

算法计算值相同，均为 0.987，1.0，0.972。RIFS 算法在 Myeloma，ALL2，ALL4 等剩余数据集上计算出的最大 Acc 值小于 CFS 算法，具体值可参见上表 4-2-1(1)。其余算法间比较数据将在下表 4-2-2 中给出。

	RIFS	CFS	PAM	RRF
RIFS	0/16/0	5/4/7	12/1/3	14/0/2
CFS	7/4/5	0/16/0	15/1/0	15/1/0
PAM	3/1/12	0/1/15	0/16/0	11/0/5
RRF	2/0/14	0/1/15	5/0/11	0/16/0

图 4-2-2 各算法 mAcc 比较

4.2.3 特征子集中特征个数比较

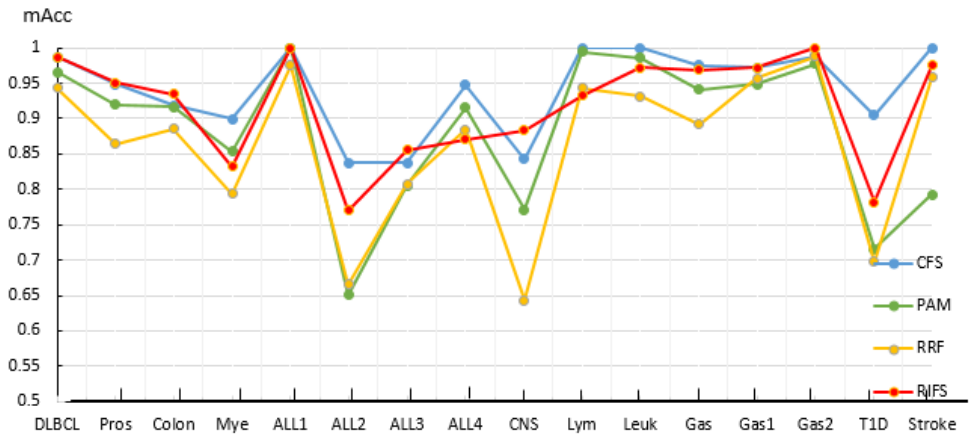
本课题中判别特征选择算法的另一项重要指标为挑选出的最优特征子集的特征个数。下表详细给出各个特征选择算法在 16 个数据集上挑选出的最优特征子集中特征数的比较。以 RIFS 算法和 PAM 算法的比较为例，PAM 算法在数据集 Prostate，ALL1，ALL2，Stroke 上挑选出的最优特征子集中特征的个数分别为 2，1，1，3，小于 RIFS 算法分别挑选出的最优特征子集特征个数出 5，8，3，4。PAM 算法在数据集 DLBCL，Colon，Adenoma 等 12 个数据集上挑选出的最优特征子集特征个数均大于 RIFS 算法挑选出的最优特征子集特征个数。特征个数与对应的特征选择算法以及相应的数据集具体值可参见上表 4-2-1(2)。其余算法间比较数据将在下表 4-2-3 中给出。

	RIFS	CFS	PAM	RRF
RIFS	0/16/0	16/0/0	12/0/4	15/0/1
CFS	7/4/5	0/16/0	15/0/1	0/0/16
PAM	4/0/12	1/0/15	0/16/0	9/0/7
RRF	1/0/15	16/0/0	7/0/9	0/16/0

表 4-2-3 各算法特征个数比较

4.3 综合结果分析

最优特征的评判准则，是最高 Acc 指标和所选择特征子集中特征个数综合作用的结果。为更直观地给出四个算法的对比情况，下图给出了四个特征选择算法在 16 个数据集上进行特征选择所得出的最大 Acc 的折线图。



CFS	141	80	53	107	103	56	43	105	39	68	39	115	300	237	166	138
PAM	53	2	14	34	1	1	2	30	20	109	20	44	29	500	48	3
RRF	12	14	16	39	10	31	34	16	29	9	29	7	16	9	28	14
RIFS	9	5	6	1	8	3	1	6	5	1	8	2	9	12	5	4

图 4-3(1) Acc 折线图

单从上述折线图可以看出，RIFS 算法具有较好的表现。除了在 Mye，ALL4，Lym 数据集上表现略逊于 PAM，RRF 算法，在其他数据集上的表现均近似或是更优于 PAM 和 RRF 算法。结合最优特征子集中特征数目来看，以 Mye 数据集为例，RIFS 算法所挑选的特征子集个数为 1，而 PAM 和 RRF 算法的特征个数分别为 14 和 16。以 CFS 算法和 RIFS 算法的对比则能更直观地看出，虽然 RIFS 在 mAcc 上的表现不如 CFS 算法，但其特征个数却相当乐观。为了更客观地比较各个特征算则算法的优劣，本文引入公式 EI=Acc-p/100,其中 EI 表示结合 Acc 指标和特征个数来判断算法表现的指标，p 指选出的最优特征子集中特征的个数。下图为四个特征选择算法在 16 个数据集上计算的 EI 折线图。

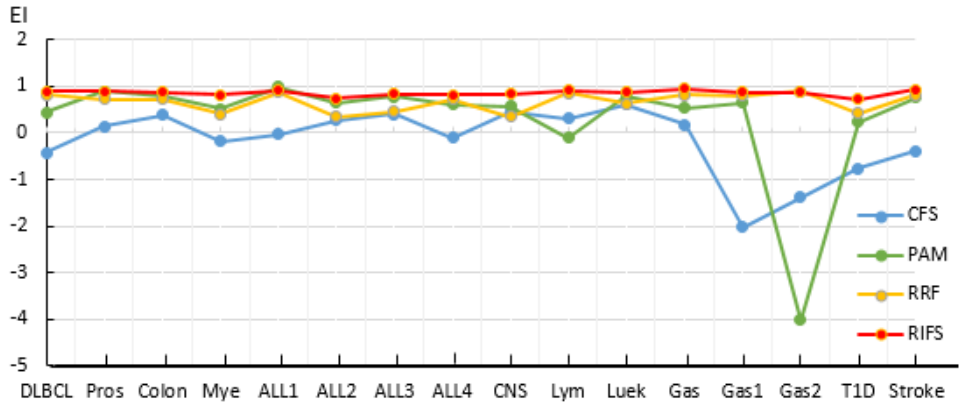


图 4-3(2) EI 折线图

通过计算，仅有 PAM 算法作用在 ALL1 和 Gas2 数据集上时，所得的 EI 值会高于 RIFS 算法，分别为 0.99 和 0.898。此时 RIFS 算法 EI 值为 0.92 和 0.88，略低于 PAM 算法。除此之外，RIFS 算法在各个数据集上的表现

均为最优。由此可见，综合 Acc 和特征个数两个指标来看，RIFS 算法类比其他算法，能挑选出更优的特征子集。

第五章 总结与展望

基于大数据的健康管理是现代医疗健康和未来医学的前沿趋势。其中基于临床数据的生物标记物的健康诊断是一个重要的研究热点，有着广阔的应用前景。癌症已然成为医学上最难克服的病症之一，如何在海量的生物数据中寻找高预测能力的肿瘤标志物组合，对癌症病症进行高效额预测评估，是一个伟大的医学研究方向。在高纬的生物特征样本中，通过一定的评估准低维的特征集合，是特征选择的过程的，特征提取作为一个困难长期存在。由于在海量数据中遍历标志物组合，求解期预测能力的计算时间成指数级爆炸性增长，现已发展了多种特征选择算法。

本文基于肿瘤标志物问题提出了基于随机重启的特征选择算法(简称 RIFS 算法)。RIFS 算法对将单个标记物预测精度，稳定性高的组合而成的特征组合也有很高的预测精度这一想法提出质疑，提出了随机重启的思想。在整个 RIFS 算法的研究过程中，加入了 3 折交叉验证，多个二元分类算法等。我们的主要研究成果如下：

（一）本文提出的 RIFS 算法首先对数据样本中各个子特征做 t-test，计算出单个特征作为肿瘤标志的预测精度，并以该预测精度为指标对各个特征进行排序。暂且保留单个标记物预测精度，稳定性高的组合而成的特征组合也有很高的预测精度的想法，这样提高算法提取特征子集的预测精度。

（二）随机生成 m 个随机数据$\{k_1, k_2 \ldots, k_m\}$，在排好序的数据样本中，分别以这 m 个随机数据为起始位置，依次往下寻找 n 个子特征，构成 m 个特征数为 n 的候选特征子集。同时，算法遍历计算由排名靠前的子特征组成的特征组合的预测精度。随机重启的思想扩大了搜索范围，排除了仅计算排名靠前的特征的弊端。

（三）RIFS 算法在计算候选特征子集的预测精度时，运用了 D-Tree，

Nbayes，KNN，SVM 四种二元分类算法。利用 3 折交叉验证提高分类的精度，尽量地减少分类误差。在计算最终，选取四种分类算法计算出的最高预测精度作为该候选特征子集的性能指标。多种分类算法的使用使得算法具有更好的计算性能，因为不同的数据集适合不同的分类算法，单一分类器并不一定完全适用于多个不同的数据集。

一个优秀的预测肿瘤标志的特征子集具有两个特点，一是具有相对高的预测精度，而是具有相对小的特征个数。在本文的实验中，对 16 个数据样本进行计算，选出各自的最优特征子集。将 RIFS 算法与现有的 CFS 算法，PAM 算法，RRF 算法相比较后，实验结果表明，本文提出的 RIFS 算法在结合最优特征子集的预测精度和其中特征个数两个指标后，比其他算法具有更好的表现性能，验证了 RIFS 算法的猜想，说明 RIFS 算法的有效性，合理性。

本文提出的算法在 IFS 算法的基础上改进,提出随机重启的思想。在测试的 16 个数据集中得到的结果相对于其它算法也表现出色。在之后的发展中，可以尝试做 5 倍交叉验证，10 倍交叉验证，提高计算的准确度。RIFS 算法在计算过程中所消耗的时间也不容忽视。在实际的应用中待筛选的特征集合容量可能更大，因此如何提升程序的性能是未来工作中需要解决的问题。这样才能将此算法投入到实际应用中，或者说应用到其它更多的领域，从而真正体现该算法的价值。

致　　谢

这篇毕业设计论文的完成，意味着我大学四年的学习旅程的结束。在这四年的学习生活中，吉林大学良好的学习氛围，严谨的科研精神引导着我的求学之路。在此，我向孕育了无数莘莘学子的母校表示诚挚的感谢，也向毕业的，未毕业的吉大人表示衷心的祝福。我特别要感谢的是我的导师周丰丰老师，这篇标志着我成长的论文正是在周丰丰老师的悉心指导下完成的。从论文选题初期的懵懂，到论文中期的困难重重，再到论文末期的浮躁，周老师一直陪伴左右，及其耐心，细心地为我提出中肯的意见。在困难面前，老师给出专业的指导和信心的鼓励。不仅在学术上，这位从前从未谋面的老师在我的生活中也为我提出了很暖心的建议，给予了我莫大的关心。不仅是老师，在整个论文的撰写过程中，同学朋友也给了我很多帮助与鼓励。在此，我向所有帮助过我，鼓励过我的老师，同学，家人表示由衷的感谢与美好的祝愿！

参考文献

[1] 王树林，王戟，陈火旺，李树涛，张波云.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

[2] 张松瑶，张绍武. 基于二元网络异步重启随机游走算法预测肺癌风险致病基因[J] .生物物理学报，31(1):33-34

[3] H Liu，R Setiono Incremental Feature Selection[J] .Applied Intelligence, 1998, volume 9(3):217-230(14)

[4] 边肇祺, 张学工. 模式识别[M]. 第 2 版. 北京: 清华大学 出版社, 2000. (Bian Z Q, Zhang X G. Pattern recognition[M]. 2nd ed. Beijing: Tsinghua University Publisher, 2000.)

[5] Manoranjan Dash, Huan Liu. Feature selection for classification[J]. Intelligent Data Analysis, 1997, 1(3):131-156.

[6] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America 1999, 96(12):6745-6750.

[7] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M,Angelo M, Reich M, Pinkus GS et al: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine 2002, 8(1):68-74.

[8] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP et al: Gene expression correlates of clinical prostate cancer behavior. Cancer cell 2002, 1(2):203-209.

[9] Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R: Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. Blood 2004, 103(7):2771-2778.

[10] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME,

Kim JY, Goumnerova LC, Black PM, Lau C et al: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002, 415(6870):436-442.

[11] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JG, Sabet H, Tran T, Yu X et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403(6769):503-511.

[12] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999,286(5439):531-537.

[13] Tian E, Zhan F, Walker R, Rasmussen E, Ma Y, Barlogie B, Shaughnessy JD, Jr.: The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. The New England journal of medicine 2003, 349(26):2483-2494.

[14] Wu YH, Grabsch H, Ivanova T, Tan IB, Murray J, Ooi CH, Wright AI, West NP,Hutchins GGA, Wu J et al: Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. Gut 2013, 62(8):1100-1111.

[15] Wang GS, Hu N, Yang HH, Wang LM, Su H, Wang CY, Clifford R, Dawsey EM, Li JM,Ding T et al: Comparison of Global Gene Expression of Gastric Cardia and Noncardia Cancers from a High-Risk Population in China. Plos One 2013, 8(5).

[16] Levy H, Wang X, Kaldunski M, Jia S, Kramer J, Pavletich SJ, Reske M, Gessel T,Yassai M, Quasney MW et al: Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. Genes Immun 2012,13(8):593-604.

[17] Krug T, Gabriel JP, Taipa R, Fonseca BV, Domingues-Montanari S, Fernandez-Cadenas I, Manso H, Gouveia LO, Sobral J, Albergaria I et al: TTC7B emerges as a novel risk factor for ischemic stroke through the

convergence of several genome-wide approaches. J Cerebr Blood F Met 2012, 32(6):1061-1072.

[18] Tibshirani R, Hastie T, Narasimhan B, Chu G: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America 2002, 99(10):6567-6572.

[19] Deng HT, Runger G: Feature Selection via Regularized Trees. Ieee Ijcnn 2012.

[20] Hall MA: Correlation-Based Feature selection for discrete and numeric class machine learning. In: Langley P, et al, eds Proc of the 17th Intl Conf Machine Learning San Francisco: Morgan Kaufmann Publishers 2000:359-366.