**McTwo论文翻译**

高通量生物OMIC技术以不断增加的速度从生物样品中生成高维数据，而传统实验中的训练样本数量由于各种困难而仍然很小。 生物医学“大数据”领域中的这种“大p，小n”范例可以至少部分地通过特征选择算法来解决，所述特征选择算法仅选择与表型显着相关的特征。特征选择是NP难问题。由于寻找全局最优解的时间要求呈指数增长，所有现有的特征选择算法都采用启发式规则来寻找局部最优解，并且它们的解在不同数据集上实现不同的性能。

该工作描述了基于最近公布的相关性测量，最大信息系数（MIC）的特征选择算法。 所提出的算法McTwo旨在相互独立地选择与表型相关的特征，并实现最近邻算法的高分类性能。基于对17个数据集的比较研究，McTwo的性能与现有算法相当或更好，所选特征的数量显着减少。McTwo选择的特征似乎与文献中的表型具有特定的生物医学相关性。

McTwo选择具有非常好的分类性能的特征子集以及小的特征编号。因此，McTwo可以代表高维生物医学数据集的互补特征选择算法。

**最大信息系数（MIC）**，**启发式算法**，**特征选择**，**滤波算法**，**包装算法**

**正文：**

由于难以收集特定样品类型和有限的可用资源，因此只能收集少量样品用于传统的生物学研究。然而，对于现代生物技术，可以为单个样品生产大量的生物医学“大数据”。这导致生物大数据中的“大p小n”范式的挑战，这是其他大数据领域中广泛使用的深度学习策略无法解决的。“大p小n”数据集通常具有数十个或最多几百个样本以及每个样本数百万或更多的特征。如果在这些样本的分类或回归建模中使用所有特征，则将引起过拟合。其中一个解决方案是特征选择或降维，它试图找到具有最佳建模性能的特征子集。

已经发布了各种特征选择算法，并且它们可以基于它们如何确定所选特征而大致分为三类。I类包装器算法通常采用现有的数据挖掘算法来评估特征子集，并对具有最优数据挖掘性能的特征子集应用启发式特征筛选规则。它倾向于消耗指数增加的时间来找到这样的特征子集。I类算法通常使用启发式规则来寻找局部最优解。微阵列预测分析（PAM）算法计算每个类标签的质心，并选择特征以将基因质心缩小到整个类质心。PAM非常适用于离群值功能。规则化随机森林（RRF）算法通过评估每个随机森林节点处的训练数据的子集上的特征来使用贪婪规则。如果新功能的信息增益不会改善所选功能的选择，则会对其进行选择。

II类滤波器算法测量每个特征或特征子集与样本标签的关联，并基于该测量对所有特征或特征子集进行排序。大多数过滤算法都会评估各个功能。对于基于特征的滤波器算法，用户可以选择确定进一步实验验证的排名靠前的特征的数量，但是没有为具有最佳建模性能的特征子集提供信息。滤波器算法不考虑特征间相关性，但其线性计算时间复杂度有时使其成为大型数据集的唯一可负担的选择。基于T检验的过滤（TRank）算法是测试两组之间特征差异的最常用方法。它估计了两组之间的差异以及数据集的变化，给出了统计显着性测量。基于威尔科克森测试的特征滤波（WRank）算法计算特征在两类样本之间如何区别的非参数分数，并且因其对异常值的鲁棒性而闻名。基于ROC图的过滤（ROCRank）算法评估特征的ROC曲线下面积（AUC）对于所研究的二元分类性能的重要程度。基于相关的特征选择（CFS）算法是基于过滤器的子集评估启发式算法，其假定良好特征子集中的特征应该彼此独立并且与样本的类标签高度相关。

III类混合算法旨在通过在不同的启发式特征选择步骤中集成包装器和滤波器策略来自动生成最佳选择的特征子集。例如，Xing，et al。提出了过滤器和包装器方法的混合，以选择高维微阵列数据集的特征子集，并且以令人满意的分类错误率优于正则化策略。

本研究提出了一种新的包装器特征选择算法McTwo，它基于两个变量之间的测量最大信息系数（MIC）。McTwo的第一步是筛选MIC与类标签相互关联的所有特征，并且只保留具有显着辨别力的那些特征以供进一步筛选。然后，McTwo采用最佳的第一搜索策略来找到具有最佳分类性能的特征子集。实验数据表明，在大多数情况下，该算法优于其他算法，并且特征数量显着减少。

**方法：**

**二元分类问题及其性能测量**

这项工作调查了二进制分类问题。 二元分类问题具有两组样本，即正（P）和负（N）集。 P = {P1，P2，...，Pn}并且N = {N1，N2，...，Nm}。正和负样本的数量也分别缩写为P=n和N=m。样本总数为s=n+m。每个样本X∈P∪N是k特征向量X = <F1(X)，F2(X)，... Fk(X)>。二元分类器f将X分配给P或N.

灵敏度（Sn），特异性（Sp）和准确度（Acc）被广泛用于衡量二元分类模型的表现。TP和FN定义分别为在阳性样本中，预测为阳性（注：预测对了，True positive）和预测为阴性（预测错了，False negative）.TN和FP的定义分别为：在阴性样本中（实际为阴性），预测为阴性（注：预测对了，True negative）和阳性（注：预测错了，False positive）。所以P=TP+FN（阳性总数=真阳性+假阴性），N=TN+FP（阴性总数=真阴性+假阳性）。灵敏度（Sn）定义为正确预测的阳性样本的比率Sn = TP/（TP + FN）=TP/P，特异性（Sp）是校正的预测阴性样本的比率Sp=TN/（TN + FP））=TN/N。模型的总体精度定义为：

另一个测量Avc定义为（Sn + Sp）/ 2，以帮助评估具有不同数量的正样本和负样本的不平衡数据集。

使用5倍内部交叉验证评估所有分类算法的总体性能测量，对于随机数发生器，使用不同种子平均超过30次运行。具有较大Acc值的二元分类算法表现更好。如果两个模型Acc值相差不大，则更简单的模型会更好，因为它在临床部署中花费更少的资源和人力。此外，更简单的模型可以避免由“大p小n”范例引起的生物医学大数据区域中的过度拟合挑战。还进行外部交叉验证以测试McTwo是否生成特征选择偏差。

所提出的特征选择算法可以为任何二元分类数据集选择特征。为了便于讨论和数据集可用性，本研究重点关注基于微阵列的基因表达谱分析数据集的分类性能比较。

**本研究中使用的生物医学数据集**

在该研究中使用17个二元分类数据集进行分类性能评估，如表1所示。两个广泛研究的数据集分别从R/Bioconductor包colonCA和golubEsets中检索结肠和白血病。从广泛研究所基因组数据分析中心下载了六个公开可用的数据集，即DLBCL,前列腺,ALL,CNS，淋巴瘤和腺瘤，该中心可从<http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi>获得。 根据表1中所述的不同表型注释，将数据集ALL进一步处理成四个二元分类数据集，即ALL1，ALL2，ALL3和ALL4。另外五个新数据集，即骨髓瘤（登录号：GDS531），胃（登录号：GSE37023） ），从NCBI Gene Expression Omnibus（GEO）数据库下载Gastric1 / Gastric2（登录号：GSE29272），T1D（登录号：GSE35725）和Stroke（登录号：GSE22255）。

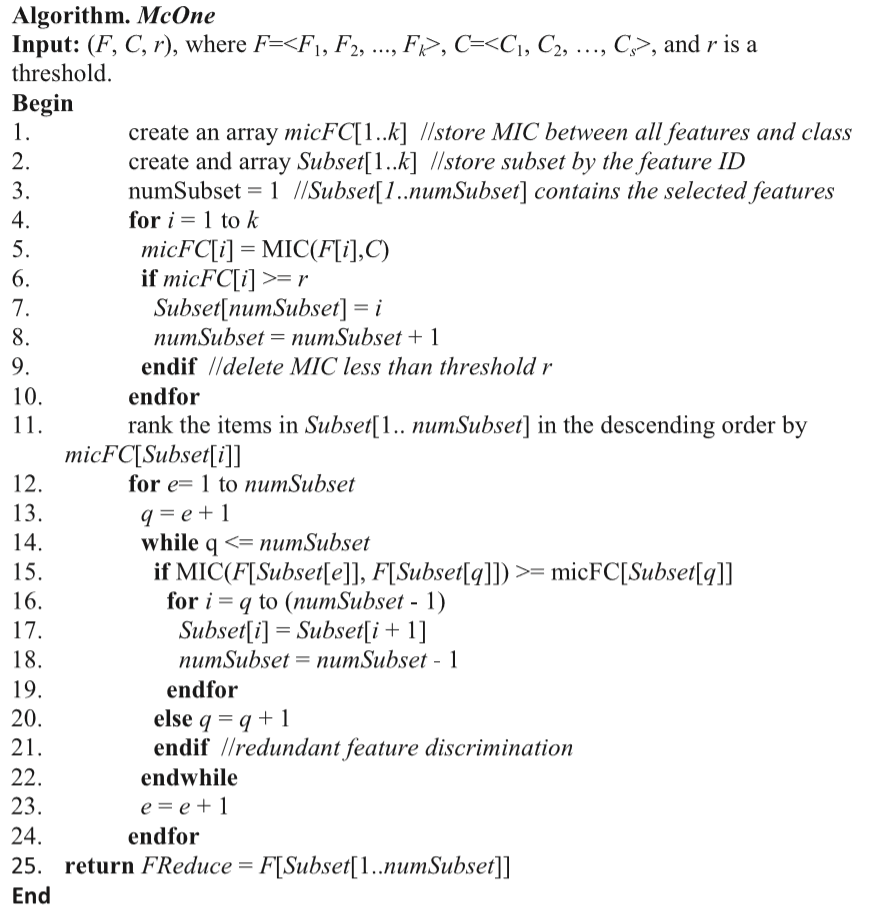
来自NCBI GEO数据库的原始数据被归一化为具有RMA算法的默认参数的基因表达矩阵，并且所有其他数据集被下载为归一化数据矩阵。

如上所述，本研究中使用的所有数据集先前由其他研究人员公布并公开可用。 因此，本研究不需要道德规范或知情同意书。

**基于最大信息系数的特征筛选（McOne）**

最大信息系数（MIC）测试两个变量之间的相关性以及它们是否具有线性或其他函数关系。测量MIC是对称的并且归一化为范围[0,1]。MIC值高表明调查变量之间的依赖性，而MIC=0表示两个独立变量之间的关系。尽管MIC对于不同的依赖类型似乎是公平的并且比其他算法（如动态切片算法和t检验）略差，但它处理数字和类别数据的能力将有助于未来应用于异构生物医学数据集。MIC的计算功能由原作者实现为R包Minerva版本1.5。 MIC的统计特征以及MIC与其他统计检验（包括Pearson校正和互信息）的综合比较可以在[14]中找到。

**算法McOne自然语言描述**



注:

irrelevancy threshold

不相关阈值

提出了基于MIC的过滤步骤McOne，以去除与表型很少关联的那些特征或者与特征子集中剩余的其他特征冗余的特征，如上述伪代码中所述。 首先，定义了许多术语。 对于给定的二元分类问题，类标记C={C1，C2，...，Cs}，Ci∈{P，N}，并且每个样本具有k个特征<F1（X），F2（X），... Fk（ X）>，其中Fj是第j个特征。

定义：信息相关特征：S={Fi | MIC（Fi，C）> r}，其中r是预设的不相关阈值。

定义：信息冗余功能：Fi是冗余的，如果存在另一个特征Fj， s.t. MIC(Fj,C)> MIC(Fi,C) and MIC(Fj,Fi)>MIC(Fi,C).

信息主导标准：如果特征Fj与候选特征子集MIC（Fj，C）中的目标变量C具有最大信息相关性并且与已经选择的特征不重复，则将保持特征Fj。

**McTwo算法**

我们采用最佳的首次搜索策略来进一步减少功能。我们的实验数据显示，McOne选择具有令人满意的分类性能的特征子集。然而，McOne可能会选择几十个甚至一百多个特征，这可能会导致某些大数据区域出现过度拟合问题，并出现“大p小n”挑战。最佳的第一搜索策略被广泛用于进一步减少小规模特征子集中的所选特征的数量。本研究使用软件R版本3.0.2中FSelector软件包版本0.19中实现的版本。

k最近邻（KNN）算法用作最佳第一搜索过程中的嵌入分类器。虽然KNN是一个非常简单的分类器，但它在快速计算和参数独立性方面的优点使得NN在包装过程中多次执行时成为理想的分类器。选择简单形式NN，即k = 1。

通过留一法（LOO）验证策略计算的平衡准确度BAcc=（Sn + Sp）/2被用作优化目标。这是因为总体准确度Acc并不总是反映不平衡数据集上分类器的合理分类性能。例如，对于100个正样本和10,000个负样本的数据集，如果分类器将任何样本预测为“负”，则其具有Acc = 10,000 /（100+10,000）=99.01％，但是Sn=0.相比之下，BAcc = （0 + 1）/2=0.5，对这样的分类器排名非常低。此外，LOO验证与参数无关，并且可以是评估分类器执行情况的客观策略。

前面提到的McTwo的第二步使用McOne的输出特征子集作为其输入，并返回通过上述过程过滤的特征。

**McTwo的时间复杂度估计**

这是对McTwo算法的时间复杂度的估计。McOne需要计算所有特征之间的MIC值，以及针对类标签的特征。设p和n分别为特征和样本的数量。因此，假设两个变量之间的MIC值是在恒定时间内计算的，那么McOne最多运行时间为O（p2 + p）。McTwo的第二步理论上需要筛选McOne过滤的所有剩余特征，最多为O（p）。 因此，McTwo的最坏情况时间复杂度是O（p2 + p）+ O（p）~O（p2 + 2p）~O（p2），这与特征选择算法FCBF相同。 但实施过滤步骤McOne是为了评估要素和类标签之间的MIC值，这通常会排除大部分功能。 然后，将显着加快对干扰MIC值的评估。 因此，在大多数情况下，实际计算时间不会达到上限O（p2）。

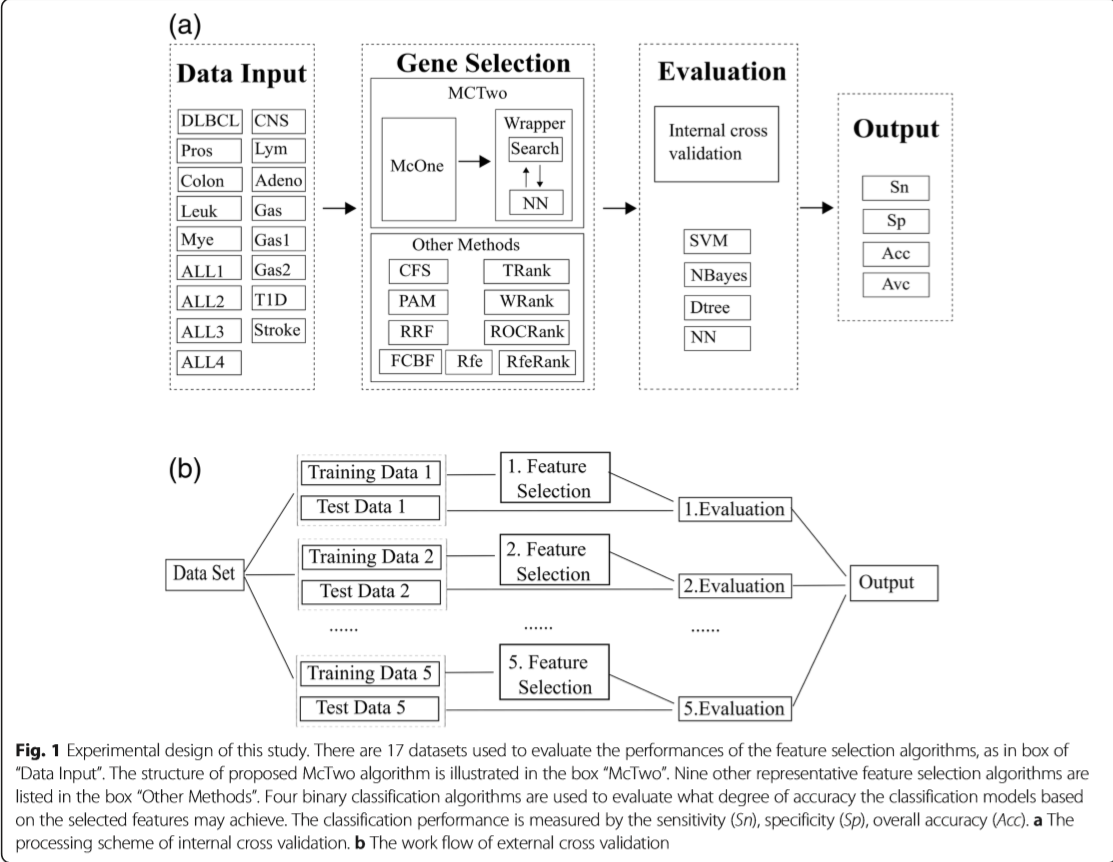
**特征选择性能的比较分析**

我们使用其他常用的特征选择算法，从分类精度和选定的特征数方面进行了一系列综合对比实验。比较是针对两种包装算法（I类），即PAM和RRF，以及三种广泛使用的滤波算法（II类），即TRank，WRank和ROCRank进行的。由于滤波算法CFS自动生成最佳选择的特征子集，因此在比较实验中将CFS与包装算法分组。

FCBF（基于快速相关的滤波器）以与McTwo类似的方式选择功能。 FCBF和McTwo之间存在两个主要差异。在第一步中，McTwo使用测量MIC来测试两个变量之间的关联，而FCBF使用基于信息增益的对称不确定性（SU）。据称MIC比其他相关测量更适合复杂数据集。在第二步中，McTwo根据嵌入式分类器NN的性能选择下一个特征，而FCBF通过评估下一个特征是否与现有特征一起构成一个近似马尔可夫毯来确定是否保留下一个特征。测量SU。进行外部交叉验证以比较两种算法的分类性能。

选择许多代表性分类算法以基于由前述特征选择算法选择的特征来构建二元分类模型。支持向量机（SVM）计算高维空间中两类样本/点之间的超平面，最大化类间距离，但最小化类内距离。朴素贝叶斯（NBayes）模型假设特征彼此独立，并选择具有最大后验概率的类标签作为预测。已知NBayes与更先进和计算密集型方法竞争，例如， SVM，在一些机器学习问题，如文本分类。决策树（DTree）由树节点上的决策规则组成，关于下一个决策步骤的路径。简单最近邻（NN）算法预测查询样本在给定距离测量中属于与其最近邻居相同的类。

整个实验过程如图1所示。



**结果和讨论**

**McTwo显着降低了McOne选择的功能**

两个数据集Gas1和T1D分别从17个数据集中选择作为癌症和心血管疾病的代表。所有其他数据集的详细结果可以在附加文件1中找到：图S1。所有17个数据集的结果将在下文中进行总结和讨论。

McTwo使用不同的分类算法实现了与McOne相似的整体精度，如图2所示。首先，McOne仅在17个NN分类算法数据集中的一个上优于McTwo，而平均而言McTwo的精度优于McOne，精度提高3.99％ 。这符合我们的预期，因为McTwo尝试最小化功能编号，同时在第二步中保持类似的整体分类准确性。唯一的例外是数据集Adeno，其中McTwo的Acc（99.8％）比McOne（100％）小0.2％。但McTwo仅使用2个功能来实现与McOne基于29feature的模型几乎相似的分类性能。平均而言，McTwo（90.99％）在总体准确度Acc方面优于McOne（86.99％），在数据集Stroke上实现了10.6％的最大改进。在17个数据集中，McTwo的DTree模型优于McOne的14个模型，例外是骨髓瘤，ALL2和CNS。 McTwo对McOne的平均改善率为3.4％。但是，对于SVM和NBayes算法，McOne的平均改进分别为3.00和4.86％，优于McTwo。这可能是因为SVM和NBayes往往对特征数量敏感，而McTwo选择的特征数量明显少于McOne，这将在以下段落中讨论。

McTwo在最佳分类模型中的表现略差于McOne，如图3所示。对于给定的特征子集，研究人员将始终选择具有最大总体准确度的分类模型。因此，四种分类算法（SVM，NBayes，DTree和NN）的最大Acc（mAcc）被用作McTwo和McOne选择的特征子集的性能测量。图3显示McTwo在mAcc中的平均损失比McOne低0.8％，但在17个数据集中有11个比McOne表现同样好或更好。对于数据集Stroke，观察到mAcc的最大差异，其中McOne的表现优于McTwo 13.4％。对于其他16个数据集，两个特征选择算法之间的平均mAcc仅有0.01％的差异。

McTwo选择的功能数量明显少于McOne，如图3所示。与McOne相比，McTwo平均只选择1/33.3的功能来实现类似的分类精度。 McTwo和McOne选择的最多功能分别为7和164。对于数据集ALL1，与McOne选择的156个特征相比，McTwo和McOne在mAcc中仅使用一个特征实现了100％的mAcc。

因此，两个步骤（即McOne和包装器）在McTwo中对于找到最佳特征子集非常重要。

**与包装器FS算法进行比较**

比较了McTwo和其他三种包装算法CFS，PAM和RRF选择的特征子集的分类性能。 McTwo选择的特征的最佳分类性能通常通过分类算法DTree和NN来实现，如图4和附加文件1：图S2所示。

McTwo比PAM和RRF算法表现要好得多，但比CFS差，如表2所示。我们使用比较三元组win/tie/lose来测量算法A表现更好的数据集数量，同样好和坏算法B通过测量最大精度mAcc，并且该三联体被定义为CT（A，B）=（win/tie/loss）。在17个数据集中，McTwo分别在12和15中表现优于PAM和RRF。但是，在14个数据集中，McTwo没有比CFS更好的mAcc。因此，与PAM和RRF相比，CFS在mAcc中表现更好，CT（CFS，PAM）=（16/1/0）和CT（CFS，RRF）=（17/0/0）。

还研究了四种包装算法的分类精度和模型复杂度之间的平衡，如图5所示。平均而言，正如我们所看到的，McTwo在mAcc中比CFS低3.5％，但是1.9和3.9％更好分别比PAM和RRF。但与CFS相比，McTwo平均只需要1 / 44.4个功能。例如，McTwo和CFS在数据集ALL1上的mAcc均达到100％，但与CFS选择的103个功能相比，McTwo仅使用一个功能。目前没有可用于对预测准确性和模型复杂性的分类模型进行评级的措施。在这里，我们为此目的定义模型复杂度和分类准确度的评估指数EI = Acc-p / 100，其中p是分类模型中使用的特征的数量。除了数据集ALL3上的PAM特征选择算法之外，与所有17个数据集上的所有其他三个包装算法相比，McTwo表现最佳。在四个数据集Colon，Mye，ALL4和Lym中，McTwo在Acc中的表现比Acc差，分别为2.4,0.2,2.1和0.1％。 McTwo和PAM为四个数据集选择的特征数字的比较表明，McTwo建议的特征数量明显减少，分别为（6对14），（7对34），（2对30）和（4对109） 。在两个数据集Gas1和Stroke上可以找到类似的观察结果，其中McTwo的表现比RRF差。 CFS和PAM还显示了不同数据集的最终选择特征数量的高波动，如图5a所示。

通常，McTwo在分类准确性和模型复杂性方面优于PAM和RRF。尽管CFS在平均测量mAcc中略微优于McTwo，但McTwo使用的功能数量明显少于CFS。使用平衡模型性能测量EI，McTwo在所有17个数据集上几乎优于所有三种包装算法，如图5b所示。

**与滤波器FS算法比较**

我们进一步比较了McTwo与三种滤波算法TRank，WRank和ROCRank的分类性能。过滤算法仅基于排名测量输出有序的特征列表。因此，为了公平比较，本研究从过滤算法排序的有序特征列表中选择前p个特征，其中p是McTwo选择的特征数。

当使用NN分类算法时，McTwo几乎在17个数据集中的15个中优于所有其他三个滤波算法。唯一的两个例外是ROCRank算法分别在数据集Pros和Adeno上使用NN在Acc中比McTwo好0.8和0.2％。与基于三种滤波算法的特征的分类性能相比，基于McTwo特征的三种其他分类算法表现得相似或更好。图6a显示使用NN的最佳McTwo模型的Acc比使用数据集Gas1上的NBayes的最佳ROCRank模型小0.3％。对于数据集T1D，基于McTwo特征的NN分类模型几乎优于所有其他分类模型。一个例外是在数据集ALL3（0.7848）上，PAM优于McTwo（0.7720），其中Acc为0.0128，如图6b所示。所有17个数据集的性能测量值Sn / Sp / Acc / Avc在附加文件1中给出：图S3。

McTwo和三个滤波器算法成对比较，结果用表3中的比较三元组进行说明。首先，McTwo在14个数据集上的表现与三个滤波器算法一样好或更好。三种滤波算法在测量mAcc中的三个不同数据集上优于McTwo。所有三种滤波算法，TRank（0.759），WRank（0.759）和ROCRank（0.749）在数据集ALL2上优于McTwo（0.716），详见附加文件1：表S2。对于所有四种算法和三种包装算法，ALL2是最难的数据集（图5和7）。 CFS在mAcc（0.837）上的表现更好，但使用了56个功能，相比之下，McTwo的功能仅为0.716，仅选择了两个功能。在所有其他情况下，滤波算法的改进mAcc值比McTwo提高不超过1.1％，如附加文件1：表S2所示。

上述数据表明，McTwo在17个数据集中的大多数数据集上的表现优于三个过滤算法，而在其他数据集上表现相似。

**特征选择算法的外部交叉验证**

进行了五次外部交叉验证，以便将McTwo与其他特征选择算法进行比较。由于CFS算法的计算要求过高，选择三个最大的数据集ALL1，Gas1和Mye进行外部交叉验证的比较研究。建议使用外部交叉验证来评估特征选择算法是否具有小数据集的选择偏差。广泛使用的特征选择算法，即基于递归特征消除（SVM-RFE）的支持向量机，可以用作滤波器或包装器模型。这些在该比较中分别表示为RfeRank和Rfe。

McTwo使用三个研究数据集上的外部交叉验证实现了令人满意且稳定的分类性能，如图8所示。如内部交叉验证的结果，McTwo在两个数据集ALL1上没有达到最佳分类性能（mAcc = 0.969）和Gas1（mAcc = 0.903），但其性能与其他算法相似。与数据集ALL1和Gas1上的包装器和滤波器算法相比，McTwo还显示出更小的变化。类似的低变化仅通过数据集ALL1上的CFS，PAM，RRF和TRank实现。数据集Mye对于所有特征选择算法都具有挑战性，其中没有一个算法实现超过0.800。所有特征选择算法都为数据集Mye产生类似的变化。 McTwo对FCBF有类似的特征筛选大纲，但在三个研究数据集上实现了更好的分类性能。这可能是因为McTwo将分类性能作为其优化目标。

还使用配对t检验评估统计显着性，以确定McTwo是否比其他每种特征选择算法表现得更好，更似或更差[48]。结果显示在附加文件1：表S3中。对于置信水平0.95，CFS和PAM在所有数据集上的统计显着优于McTwo。而McTwo的表现与所有其他包装算法相似或略好。除数据集Mye外，McTwo在所有数据集上的性能均优于所有调查过滤算法。当选择稍微更严格的置信水平0.99时，除了数据集Mye之外，没有任何特征选择算法比McTwo表现更好。算法CFS和PAM的性能优于McTwo，具有统计显着性。然而，McTwo使用的功能仅为CFS和PAM的1/3，以实现类似或稍差的分类性能。

**最佳包装器功能并不总是通过过滤器算法排名靠前**

如图6所示，最好的McTwo模型与三种滤波器特征选择算法的表现相似或更好，但是所选择的特征并不总是由滤波器算法评估的排名靠前的特征。表4总结了McTwo选择的Gastric1和6个T1D特征中的每一个特征如何按三种滤波算法排序。

除了探针组216381\_x\_at分别由TRank，WRank和ROCRank分别为9,9和1之外，McTwo选择的所有其他特征的排名都低于25，如表4所示。例如，数据集Gastric1甚至具有在分类模型中选择831排名特征，总体准确度为95.35％。数据集T1D具有54,675个特征，基于McTwo的NN分类模型在整体精度方面优于所有其他模型，如图6b所示。但是这个最佳模型使用了ROCRank在54,675功能中排名52,455的功能。基于过滤算法的广泛使用的特征选择策略是选择排名前K的特征，其中K通常通过反复试验来确定。因此，任何过滤算法通常都会删除这些低排名的功能。

我们的数据表明，最佳分类模型可能会使用一些按滤波算法排名较低的特征。这是合理的，因为过滤算法独立地评估每个特征与类标签的关联，并且排名最高的特征的组合不一定导致具有高总体准确度的分类模型。例如，与排名最高的特征线性相关的特征也将高度排名。然而，这些线性相关的高排名特征的组合将不会改善基于排名最高的特征的分类模型。独立于排名靠前的特征的排名较低的特征可以导致更好的分类模型。

**McTwo所选特征的生物学推论**

虽然McTwo选择的大部分功能都被滤波器算法排名较低，但许多功能在疾病发生和发展方面都具有已知的作用。例如，已知Gastric1特征中的两个2​​16381\_x\_at和218595\_s\_at与胃癌相关，如表4所示。基因AKR7A3的探针组216381\_x\_at（醛酮还原酶家族7，成员A3）参与生物过程。细胞醛代谢和氧化还原。一项独立研究观察了胃癌与对照样本之间的差异转录水平。 HEATR1基因（含有1的HEAT重复序列）的探针组218595\_s\_at可以预防幽门螺杆菌感染的胃上皮细胞中的细胞凋亡并诱导胃癌。

另外两个探针组209902\_at和205523\_at与许多癌症类型广泛相关，但是它们与胃癌的关联需要进一步研究。探针组209902\_at的基因ATR（共济失调性毛细血管扩张症和Rad3相关;类似于共济失调性毛细血管扩张症和Rad3相关蛋白）是丝氨酸/苏氨酸蛋白激酶。 ATR作为DNA损伤传感器，当暴露于IR（电离辐射）和UV（紫外线）等有害影响时，激活检查点信号，如BRCA1，CHEK1，MCM2，RAD17，RPA2和p53 / TP53。这些条件可导致阻断DNA复制和有丝分裂，并促进DNA修复和细胞凋亡。 ATR与各种类型的癌症有关，例如食道腺癌，口咽癌，子宫内膜样子宫内膜癌，乳腺癌，卵巢癌等。基因HAPLN1（透明质酸和蛋白多糖连接蛋白1）的探针组205523\_at可以将蛋白多糖单体和透明质酸的聚合物保持在软骨基质中。 HAPLN1涉及从细胞粘附到生物粘附的生物过程。已知HAPLN1与许多癌症类型相关，例如食道腺癌，乳腺癌，结肠直肠癌等。最近的一项研究表明，HAPLN1及其SP-IgV结构域的过度表达改善了恶性胸膜间皮瘤的致瘤性。因此，HAPLN1可能与癌症治疗有关。

McTwo选择的六种T1D特征之一，即208031\_s\_at，也已知与糖尿病的发展密切相关。基因RFX2的探针组208031\_s\_at（调节因子X，2（影响HLA II类表达））是转录因子。转录激活因子rfx2可以与IL-5α受体α基因启动子中的DNA结合。 RFX2参与转录的生物学过程，转录的调节和RNA代谢的调节。已经证明RFX2在DREAM（用雷米普利和罗格列酮药物治疗的糖尿病减少评估）研究中在糖尿病的发展中起重要作用。

**结论**

这项研究描述了一种新的基于MIC的包装器特征选择算法McTwo。 与现有的特征选择算法相比，McTwo实现了更好或类似的分类性能，并且与其他包装算法相比，推荐了更少的特征。 与其他滤波器算法相比，使用相同数量的功能，McTwo也实现了更好或相似的性能。 McTwo选择的特征可能会导致有趣的生物学假设进行进一步的实验研究。

可能需要参考的网站：https://blog.csdn.net/FontThrone/article/details/85227239