



# 一种预测个体肿瘤的抗癌药物反应分类 计算模型及其应用\*

李少达 李玉双\*\*

(燕山大学理学院, 秦皇岛 066004)

**摘要** 目的 不同患者对同一抗癌药物的反应可能不同, 了解患者之间对抗癌药物的反应差异对癌症精准医疗具有重大参考价值。**方法** 高通量测序数据为构建抗癌药物反应分类预测模型提供了强大的数据支撑。针对两大经典数据集癌症细胞百科全书 (CCLE) 和癌症药物敏感性基因组学数据集 (GDSC), 本文提出了基于最大相关最小冗余 (mRMR) 算法和支持向量机 (SVM) 的计算模型 mRMR-SVM。利用基因表达数据, 通过方差排序和 mRMR 算法提取特征基因, 借助 SVM 实现抗癌药物对细胞系的“敏感-抑制”二分类预测。**结果** 对于 CCLE 中的 22 种药物, mRMR-SVM 的平均准确率为 0.904; 对于 GDSC 中的 11 种药物, 平均准确率为 0.851。**结论** mRMR-SVM 不仅在预测性能方面优于传统的支持向量机、随机森林、深度反应森林、深度神经网络和细胞系-药物复杂网络模型, 而且具有良好的泛化能力, 对于三类特定组织的抗癌药物反应分类预测也取得了令人满意的结果。此外, mRMR-SVM 可以识别与癌症发生发展密切相关的生物标志物。

**关键词** 抗癌药物反应, 最大相关最小冗余, 支持向量机, 敏感, 抑制

**中图分类号** Q5

**DOI:** 10.16476/j.pibb.2021.0082

癌症的异质性和遗传多样性, 导致同种癌症的患者即使采用相同的治疗方法, 也有可能得到不同的疗效<sup>[1-3]</sup>。从患者的角度, 更希望了解给定药物是否有效。针对特定癌症类型, 如何在分子水平上探索癌细胞系对抗癌药物的反应已成为精准医疗的研究热点之一<sup>[4]</sup>。基因组学的快速发展及人类基因组计划的顺利实施, 诞生了海量的生物学数据, 为在分子水平上预测抗癌药物临床反应提供了良好的数据基础<sup>[5]</sup>。特别值得一提的是, 2012 年《自然》(Nature) 杂志发表了两项系统的大规模研究, 癌症基因组计划 (CGP)<sup>[6]</sup> 和癌细胞百科全书 (CCLE)<sup>[7]</sup>, 研究中所涉及到的细胞系几乎涵盖所有常见的癌症类型, 使得完全以数据驱动、计算建模的方式自动识别生物标志物, 系统解析抗癌药物反应与癌症细胞系基因谱之间的关系成为可能。

研究人员借助这些数据集开发抗癌药物反应预测计算模型<sup>[8-11]</sup> 的主要思想有两种, 一种是基于

核方法预测药物敏感性, 其中最具有代表性的模型之一为支持向量机 (SVM)。如 Hejase 等<sup>[12]</sup> 基于美国国家癌症研究中心 (NCI) 数据, 应用非线性 SVM 成功预测了药物化合物对乳腺癌细胞的影响; Wang 等<sup>[13]</sup> 基于 CCLE 数据集, 利用基因突变、拷贝数变异和基因表达等数据, 通过组合 SVM 模型对三类特定组织下的细胞系进行了敏感性分类。另一种是基于特征提取方法预测抗癌药物反应。如最近 Su 等<sup>[14]</sup> 融合基因表达和拷贝数变异, 构建了两类深度反应森林 (Deep-Resp-Forest) 模型 MIMGS1 和 MIMGS2, 成功预测抗癌药物对细胞系的敏感或抑制。

上述模型大大推进了抗癌药物反应预测的研究

\* 河北省自然科学基金 (A2020203021) 和河北省引进留学人员自然科学基金 (C20200365) 资助项目。

\*\* 通讯联系人。

Tel: 15232356135, E-mail: yushuangli@ysu.edu.cn

收稿日期: 2021-04-01, 接受日期: 2021-05-10

进程,但在模型预测性能、应用范围等方面仍有可探索的空间。受以上工作启发,本文聚焦抗癌药物敏感-抑制二分类问题,构建了mRMR-SVM模型,从细胞系的基因表达数据出发,利用“最大相关最小冗余”算法<sup>[15]</sup>(mRMR)提取特征基因,借助SVM进行分类预测,不仅降低了时间运行成本,而且提升了模型的预测性能和生物可解释性。

## 1 数据来源和数据处理

从CCLE数据库(<http://www.broadinstitute.org/ccle>)下载了1 036个癌症细胞系的53 619个基因表达信息,以及504个细胞系对24种药物的敏感性数据(敏感性指标为activity area)。进一步选出462个既有基因表达又有药物敏感性数据的细胞系,并用z-score方法标准化敏感性数据。依据文献<sup>[14]</sup>,如果细胞系对药物的敏感性值(标准化后的敏感性值)大于0.8,定义为“细胞系对药物是敏感的”,如果小于-0.8,定义为“细胞系对药物是抑制的”,其余数据定义为冗余数据,不参与实验。在此定义下,有2种药物对应的细胞系很少,故在实验中舍去,保留其余22种药物进行分类预测,其对应细胞系的数量范围是93~215。

为验证模型的泛化性能,选取另一数据库进行实验。从癌症药物敏感性基因组学数据集(GDSC)(<https://www.cancerrxgene.org>)下载了789个癌细胞系的12 072个基因表达信息,655个癌细胞系对140种药物的敏感性数据(敏感性指标为 $IC_{50}$ )。采用与CCLE相同的处理方式,选取11种药物进行分类预测,其对应细胞系的数量范围是76~179。

## 2 mRMR-SVM模型

本文首先利用mRMR提取特征基因,然后构建SVM预测抗癌药物反应分类及识别生物标志物。具体流程如图1所示。

### 2.1 特征基因的选取

由于CCLE的基因数量大,许多基因的表达值差别不明显,为降低模型运行成本,本文先计算每个基因在所有细胞系下的表达方差,再将基因按方差从高到低的顺序进行排序,选取表达差异较大的前10 000个基因作为候选特征基因。然后利用mRMR算法提取得分最高的基因集合作为最终的

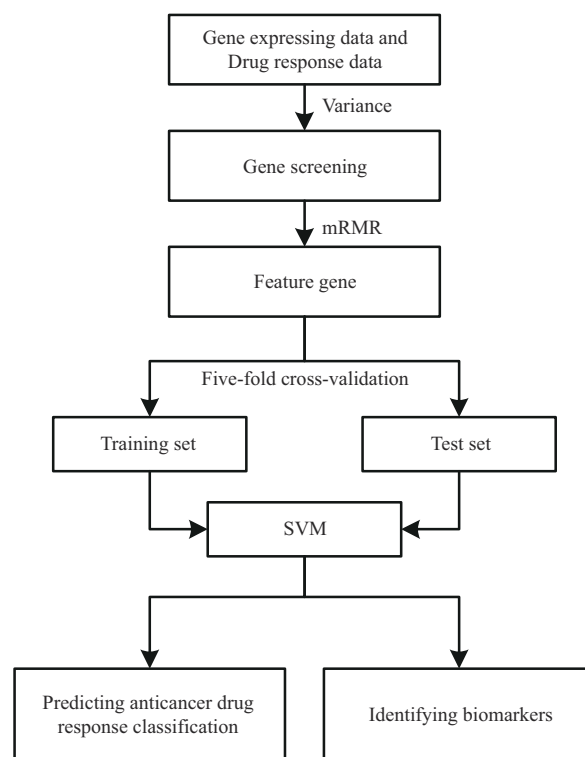


Fig. 1 The flowchart of mRMR-SVM

特征基因集。具体定义如下: 设 $x, y$ 为随机变量,  $p(x), p(y), p(x, y)$ 为概率密度函数, 则 $x$ 和 $y$ 之间的互信息为:  $I(x; y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy$ 。设 $S$ 为基因表达向量的集合(为方便计算, 本文选取 $|S| = 500$ ),  $c$ 为给定药物在所有细胞系下观测到的“敏感-抑制”类别向量: 如果第 $j$ 个细胞系对给定药物是敏感的, 其分量 $c_j = 1$ ; 如果第 $j$ 个细胞系对给定药物是抑制的,  $c_j = -1$ 。定义 $S$ 与 $c$ 的相关度 $D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c)$ , 即 $S$ 中的基因表达向量与类别向量 $c$ 之间的所有互信息值的平均值, 这里 $x_i$ 为 $S$ 中第 $i$ 个基因在所有细胞系下的表达向量。定义 $S$ 的冗余度 $R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j)$ , 即 $S$ 中基因表达向量之间所有互信息值的平均值。定义 $S$ 的得分 $\Phi(S) = D(S, c) - R(S)$ , 则 $mRMR = \max_S \Phi(S)$ 。

由于GDSC的基因数量相对较小, 故在实验中不需进行基因初筛, 直接利用mRMR算法提取特征基因。

2.2 mRMR-SVM的构建

SVM解决线性不可分问题的主要思想是: 将原始低维线性不可分的分类空间映射到高维的特征空间, 只要映射的空间维数足够高, 则原始空间将转换为一个新的线性可分空间。通过在线性可分空间建立一个最优的决策超平面, 使得距离分类平面两侧最近的训练样本之间距离最大, 将线性不可分

的数据转化为线性可分。本文构建的SVM包含两个参数, 即低维空间映射到高维空间的核函数以及惩罚因子C。

利用mRMR选取的500个特征基因的表达数据和观测到的反应分类标签来训练SVM, 具体的mRMR-SVM构建过程如图2所示。

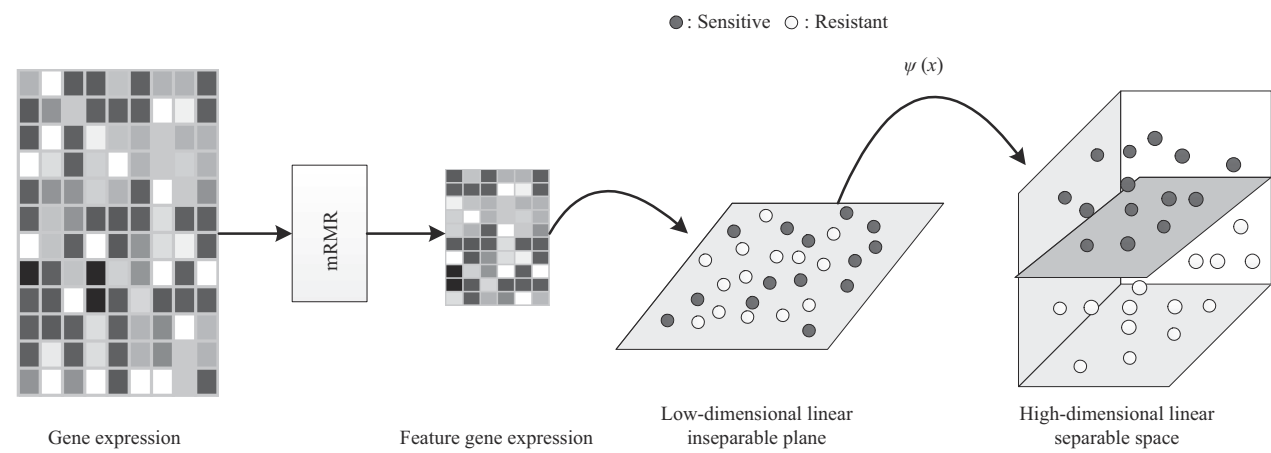


Fig. 2 The construction of mRMR-SVM

采用交叉验证确定模型参数及评估模型性能。首先, 将数据集随机分为90%的训练集和10%的测试集; 为防止模型过拟合, 再将训练集随机平均划分为5份, 分别用其中4份作为训练集, 1份作为验证集训练模型超参数; 最后, 用测试集进行模型性能评估。上述过程重复执行5次, 测试集分类结果的均值为最终预测结果。本文所用编程语言为Python, 代码见本文网络版附件。

2.3 模型的评价指标

本文采用ACC、AUC、precision、recall和F1 5个指标来评价模型的预测性能。

ACC为模型的预测准确率, 具体定义为:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

其中, TP为预测是敏感而实际也是敏感

的细胞系数数量。AUC (area under curve) 为利用预测结果所绘制的ROC曲线下面积, ROC曲线的纵坐标为TPR (true positive rate), 横坐标为FPR (false positive rate), 这里TPR为真阳率, FPR为假阳率。

precision为精确率, 反应了模型预测为敏感的细胞系的预测准确率, 定义为:

$$precision = \frac{TP}{TP + FP}$$

recall为召回率, 反应了模型对敏感细胞系的预测准确率, 定义为:

$$recall = \frac{TP}{TP + FN}$$

F1得分是综合precision和recall给出的平均定义, 其值越大, 说明模型预测性能越好。定义为:

$$F1 = 2 \frac{precision \times recall}{precision + recall}$$

### 3 抗癌药物反应分类预测结果

模型训练的最优核函数为 linear。对于惩罚因子  $C$ ，本文采用了两种确定方法，第一种直接使用模型默认的参数 1，第二种针对每种药物单独调整选出最优参数。

#### 3.1 基于CCLE数据集的预测结果分析与比较

针对 CCLE 数据集的 22 种药物，选取  $C$  为默认值 1 的 mRMR-SVM 预测结果（表 1）：22 种药物的平均  $ACC$ 、 $AUC$ 、 $precision$ 、 $recall$ 、 $F1$  分别为 0.897、0.966、0.898、0.892、0.888。单独调  $C$  的预测结果（表 2）：平均  $ACC$ 、 $AUC$ 、 $precision$ 、 $recall$ 、 $F1$  分别为 0.904、0.969、0.905、0.898、0.895。从预测结果可以看出，单独调  $C$  的模型预测结果更理想。

为了阐释 mRMR 算法提取的 500 个特征基因对抗癌药物反应分类预测的影响，一方面，利用

**Table 1** Classification result of mRMR-SVM on CCLE data set ( $C=1$ )

| Drugs        | $ACC$ | $AUC$ | $Precision$ | $Recall$ | $F1$  | $C$ |
|--------------|-------|-------|-------------|----------|-------|-----|
| 17-AAG       | 0.891 | 0.959 | 0.890       | 0.907    | 0.894 | 1   |
| AEW541       | 0.924 | 0.973 | 0.905       | 0.936    | 0.916 | 1   |
| AZD0530      | 0.872 | 0.942 | 0.869       | 0.883    | 0.869 | 1   |
| AZD6244      | 0.958 | 0.990 | 0.964       | 0.945    | 0.952 | 1   |
| Erlotinib    | 0.905 | 0.976 | 0.902       | 0.887    | 0.887 | 1   |
| Irinotecan   | 0.945 | 0.986 | 0.929       | 0.955    | 0.939 | 1   |
| LBW242       | 0.923 | 0.978 | 0.935       | 0.905    | 0.915 | 1   |
| Lapatinib    | 0.932 | 0.987 | 0.969       | 0.904    | 0.932 | 1   |
| Nutlin-3     | 0.891 | 0.975 | 0.860       | 0.923    | 0.881 | 1   |
| PD-0325901   | 0.88  | 0.952 | 0.886       | 0.856    | 0.863 | 1   |
| PD-0332991   | 0.877 | 0.965 | 0.877       | 0.868    | 0.857 | 1   |
| PF2341066    | 0.911 | 0.981 | 0.874       | 0.930    | 0.896 | 1   |
| PHA-665752   | 0.904 | 0.970 | 0.958       | 0.819    | 0.874 | 1   |
| PLX4720      | 0.944 | 0.998 | 0.924       | 0.994    | 0.952 | 1   |
| Paclitaxel   | 0.975 | 0.966 | 0.868       | 0.904    | 0.881 | 1   |
| Panobinostat | 0.846 | 0.918 | 0.811       | 0.879    | 0.836 | 1   |
| RAF265       | 0.86  | 0.946 | 0.895       | 0.807    | 0.842 | 1   |
| Sorafenib    | 0.93  | 0.992 | 0.950       | 0.896    | 0.917 | 1   |
| TAE684       | 0.904 | 0.967 | 0.902       | 0.898    | 0.894 | 1   |
| TKI258       | 0.895 | 0.967 | 0.905       | 0.879    | 0.887 | 1   |
| Topotecan    | 0.778 | 0.894 | 0.790       | 0.776    | 0.774 | 1   |
| ZD-6474      | 0.891 | 0.970 | 0.888       | 0.876    | 0.872 | 1   |
| Average      | 0.897 | 0.966 | 0.898       | 0.892    | 0.888 | /   |

**Table 2** Classification result of mRMR-SVM on CCLE data set ( $C$  separately adjusted)

| Drugs        | $ACC$ | $AUC$ | $Precision$ | $Recall$ | $F1$  | $C$ |
|--------------|-------|-------|-------------|----------|-------|-----|
| 17-AAG       | 0.891 | 0.959 | 0.890       | 0.907    | 0.894 | 1   |
| AEW541       | 0.924 | 0.973 | 0.905       | 0.936    | 0.916 | 1   |
| AZD0530      | 0.880 | 0.957 | 0.873       | 0.887    | 0.874 | 0.2 |
| AZD6244      | 0.964 | 0.992 | 0.969       | 0.953    | 0.959 | 0.2 |
| Erlotinib    | 0.915 | 0.979 | 0.935       | 0.874    | 0.896 | 0.3 |
| Irinotecan   | 0.945 | 0.986 | 0.929       | 0.955    | 0.939 | 0.6 |
| LBW242       | 0.923 | 0.978 | 0.935       | 0.905    | 0.915 | 0.3 |
| Lapatinib    | 0.950 | 0.990 | 0.973       | 0.932    | 0.950 | 0.2 |
| Nutlin-3     | 0.891 | 0.975 | 0.860       | 0.923    | 0.881 | 1   |
| PD-0325901   | 0.886 | 0.960 | 0.889       | 0.871    | 0.875 | 0.2 |
| PD-0332991   | 0.888 | 0.970 | 0.871       | 0.900    | 0.873 | 0.1 |
| PF2341066    | 0.932 | 0.984 | 0.890       | 0.954    | 0.918 | 0.1 |
| PHA-665752   | 0.906 | 0.969 | 0.955       | 0.824    | 0.877 | 0.4 |
| PLX4720      | 0.944 | 0.998 | 0.924       | 0.994    | 0.952 | 0.8 |
| Paclitaxel   | 0.875 | 0.968 | 0.858       | 0.923    | 0.885 | 0.3 |
| Panobinostat | 0.851 | 0.917 | 0.808       | 0.890    | 0.840 | 0.4 |
| RAF265       | 0.871 | 0.956 | 0.916       | 0.816    | 0.857 | 0.1 |
| Sorafenib    | 0.940 | 0.992 | 0.963       | 0.910    | 0.931 | 0.2 |
| TAE684       | 0.904 | 0.967 | 0.902       | 0.898    | 0.894 | 1   |
| TKI258       | 0.901 | 0.968 | 0.928       | 0.868    | 0.892 | 0.1 |
| Topotecan    | 0.806 | 0.906 | 0.835       | 0.768    | 0.792 | 0.1 |
| ZD-6474      | 0.898 | 0.971 | 0.907       | 0.869    | 0.879 | 0.2 |
| Average      | 0.904 | 0.969 | 0.905       | 0.898    | 0.895 | /   |

mRMR 算法提取的 500 个特征基因训练深度神经网络 (DNN)，简称 mRMR-DNN。使用网格搜索法调参，最终确定 mRMR-DNN 包含 3 个隐藏层，每个隐藏层的神经元个数分别为 60、30 和 30，层与层之间的激活函数分别为 tanh、rectifier 和 linear，输出层的激活函数为 softmax。另一方面，从经过方差筛选出的 10 000 个基因中随机挑选 500 个基因，训练 SVM 和随机森林 (RF)，两种模型均采用网格搜索法调参，SVM 的最终参数  $C=0.1$ ，RF 的参数（决策树的个数）为 80。

在共同讨论的 14 种抗癌药物反应分类预测中，mRMR-SVM（单独调参）的平均  $ACC$  为 0.911，mRMR-DNN 为 0.899，SVM 为 0.525，RF 为 0.526，文献[14]中的 MIMGS1 和 MIMGS2 分别为 0.858、0.850，说明 mRMR 算法提取的 500 个特征基因对抗癌药物反应分类预测至关重要。图 3 从整体上展示了以上 6 种模型的预测性能，mRMR-SVM 明显

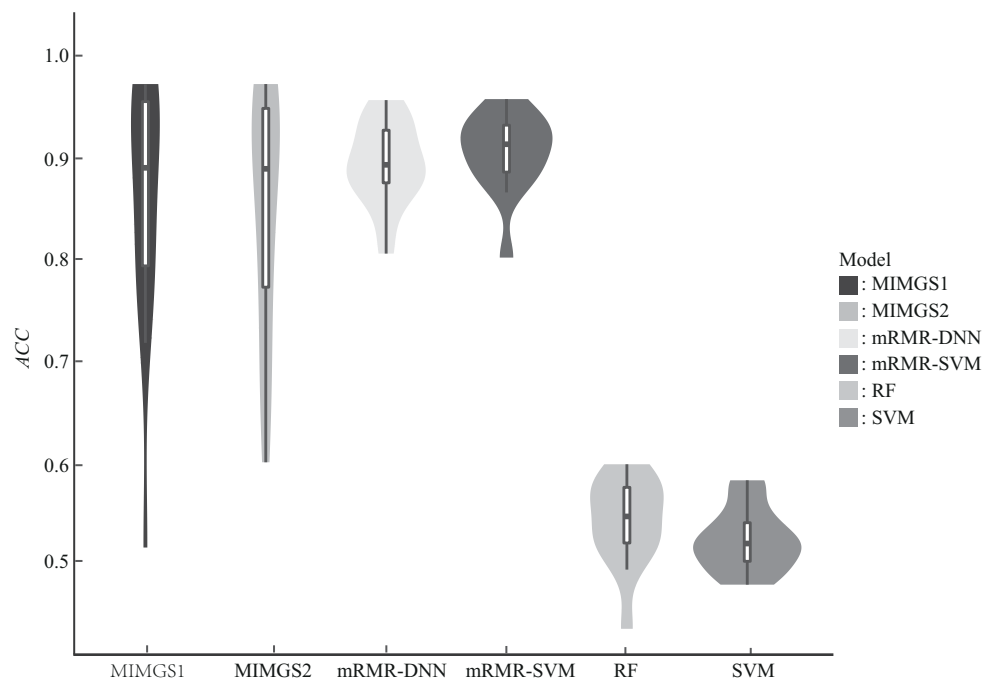


Fig. 3 The classification accuracy of six models on CCLE data set

优于其他5种模型。此外,与文献[16]中的CDCN模型进行比较,在共同讨论的22种抗癌药物的反应分类预测中,mRMR-SVM的平均ACC为0.904,高于CDCN(0.566)。

3.2 基于GDSC数据集的预测结果分析与比较

针对GDSC数据集中的11种药物,表3展示了mRMR-SVM在C为默认值1时的预测结果:平均ACC、AUC、precision、recall、F1分别为0.839、0.909、0.850、0.840、0.834。单独调C的预测性能

Table 3 Classification result of mRMR-SVM on GDSC data set (C=1)

| Drugs               | ACC   | AUC   | Precision | Recall | F1    | C |
|---------------------|-------|-------|-----------|--------|-------|---|
| Erlotinib           | 0.790 | 0.880 | 0.797     | 0.796  | 0.787 | 1 |
| Sunitinib           | 0.843 | 0.908 | 0.863     | 0.867  | 0.850 | 1 |
| AZD628              | 0.926 | 0.986 | 0.934     | 0.940  | 0.928 | 1 |
| Imatinib            | 0.930 | 0.969 | 0.928     | 0.950  | 0.928 | 1 |
| S-Trityl-L-cysteine | 0.760 | 0.850 | 0.758     | 0.775  | 0.751 | 1 |
| Z-LLNle-CHO         | 0.819 | 0.858 | 0.818     | 0.789  | 0.792 | 1 |
| GNF-2               | 0.876 | 0.940 | 0.872     | 0.885  | 0.865 | 1 |
| A-770041            | 0.797 | 0.870 | 0.828     | 0.810  | 0.808 | 1 |
| WH-4-023            | 0.870 | 0.940 | 0.883     | 0.830  | 0.850 | 1 |
| BMS-536924          | 0.792 | 0.886 | 0.828     | 0.770  | 0.787 | 1 |
| Pyrimethamine       | 0.822 | 0.911 | 0.846     | 0.825  | 0.827 | 1 |
| Average             | 0.839 | 0.909 | 0.850     | 0.840  | 0.834 | / |

进一步提升(表4),平均ACC、AUC、precision、recall、F1分别为0.851、0.917、0.865、0.848、0.845。

Table 4 Classification result of mRMR-SVM on GDSC data set (C separately adjusted)

| Drugs               | ACC   | AUC   | Precision | Recall | F1    | C   |
|---------------------|-------|-------|-----------|--------|-------|-----|
| Erlotinib           | 0.797 | 0.897 | 0.797     | 0.816  | 0.799 | 0.1 |
| Sunitinib           | 0.851 | 0.913 | 0.858     | 0.894  | 0.864 | 0.1 |
| AZD628              | 0.926 | 0.986 | 0.934     | 0.940  | 0.928 | 0.3 |
| Imatinib            | 0.935 | 0.969 | 0.928     | 0.970  | 0.941 | 0.1 |
| S-Trityl-L-cysteine | 0.803 | 0.869 | 0.811     | 0.795  | 0.785 | 0.1 |
| Z-LLNle-CHO         | 0.827 | 0.873 | 0.835     | 0.784  | 0.800 | 0.1 |
| GNF-2               | 0.876 | 0.940 | 0.872     | 0.885  | 0.865 | 0.2 |
| A-770041            | 0.819 | 0.888 | 0.848     | 0.819  | 0.825 | 0.1 |
| WH-4-023            | 0.87  | 0.940 | 0.883     | 0.830  | 0.850 | 0.3 |
| BMS-536924          | 0.827 | 0.897 | 0.888     | 0.772  | 0.812 | 0.1 |
| Pyrimethamine       | 0.829 | 0.915 | 0.859     | 0.823  | 0.831 | 0.2 |
| Average             | 0.851 | 0.917 | 0.865     | 0.848  | 0.845 | /   |

对于GDSC数据集,本文同样训练了mRMR-DNN:包含3个隐藏层,每层的神经元个数分别为63、78和86,层与层之间的激活函数分别为rectifier、linear和tanh,输出层的激活函数为softmax。同样地,从GDSC数据集12 072个基因中随机挑选500个作为特征基因,训练了SVM

( $C=0.9$ ) 和 RF (决策树个数为 95)。针对共同讨论的 11 种药物, mRMR-SVM (单独调  $C$ ) 的平均  $ACC$  为 0.851, mRMR-DNN 为 0.817, SVM 为 0.652, RF 为 0.640, MIMGS1 为 0.805, MIMGS2 为 0.815。图 4 从整体上展示了 6 种模型的预测结

果, mRMR-SVM 的预测性能明显优于其他 5 种模型。此外, 针对共同讨论的 7 种药物, mRMR-SVM 的平均  $ACC$  为 0.861, 高于文献[16]的 CDCN 模型 (0.630)。

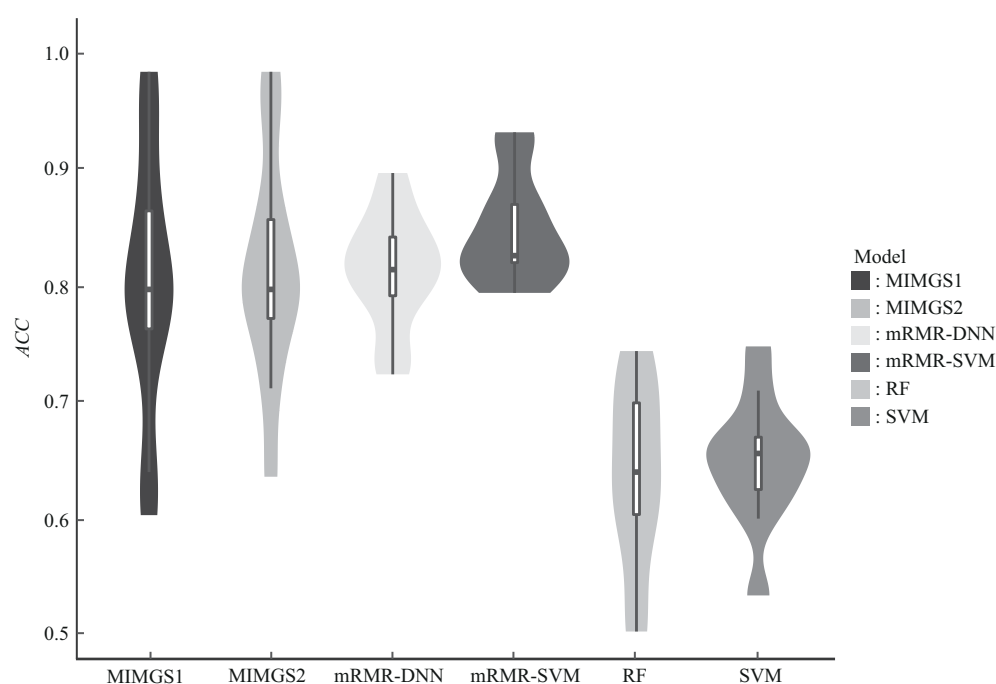


Fig. 4 The classification accuracy of six models on GDSC data set

### 3.3 基于三类特定组织的预测结果分析与比较

为进一步验证 mRMR-SVM 的泛化能力, 受文献[13]的启发, 对 CCLE 数据集中三类特定组织下的细胞系, 包括造血和淋巴组织 (包含 71 个细胞系)、皮肤组织 (包含 40 个细胞系)、肺组织 (包含 94 个细胞系), 针对 22 种抗癌药物进行反应分类预测。考虑到模型的泛化性, 参数  $C$  取默认值 1。五次五折交叉验证得到的平均预测结果如表 5 所示。三类特定组织的平均  $AUC$  依次达到了 0.973、0.981、0.965, 均优于文献[13]中基于基因表达、拷贝数变异、基因突变等多类数据融合的 SVM (其平均  $AUC$  依次为 0.81、0.82、0.83)。该实验表

明, mRMR-SVM 对于小样本数据集同样具有很好的预测能力。

## 4 生物标志物的识别

mRMR-SVM 能够识别出许多与癌症发生、发展密切相关的重要基因, 为抗癌药物生物标志物的筛选提供理论参考。如在抗癌药物 17-AAG 的特征基因中排序第二的 TP73-AS1, 已被证实在大多数肿瘤中高表达, 在乳腺癌、胃癌和肝癌等肿瘤中发挥促癌基因作用, 在膀胱癌中低表达并发挥抑癌基因作用<sup>[17]</sup>。PARK7 在 4 种药物 17-AAG、Nutlin-3、Panobinostat 和 RAF265 中均被选为 Top 基因。事实上, PARK7 已被确定为各种癌症的发病机制和生存的高危因素, 它增强了肿瘤的起始、增殖、转移和复发, 以及对化疗的抵抗力<sup>[18]</sup>。有文献发现, PQLC2 的上调对于体外和体内胃癌的发展至关重要, 靶向 PQLC2 是胃癌治疗的有效策略<sup>[19]</sup>。本文验证了 PQLC2 不仅是具有抗胃癌活性药物

Table 5 Average classification result of mRMR-SVM on three kinds of tissues

| Tissue                      | ACC   | AUC   | Precision | Recall | F1    |
|-----------------------------|-------|-------|-----------|--------|-------|
| Haematopoietic and lymphoid | 0.901 | 0.973 | 0.94      | 0.916  | 0.925 |
| Skin                        | 0.868 | 0.981 | 0.681     | 0.879  | 0.741 |
| Lung                        | 0.895 | 0.965 | 0.823     | 0.928  | 0.863 |

ZD-6474 的 Top 基因, 而且在另外 3 种抗癌药物 PD-0332991、TAE684 和 Topotecan 的特征基因排序中也位居前 5。IFI6 是一个能被 I 型干扰素诱导上调的干扰素刺激基因, 在多种恶性肿瘤中高表达, 能够抵抗细胞凋亡, 对肿瘤的放化疗效果有一定影响<sup>[20]</sup>。文中 IFI6 在 5 种药物 17-AAG、Erlotinib、TKI258、ZD-6474 和 AZD0530 的特征基因排序中均位居前 10, 与已有结果一致。

## 5 结 论

本文提出的 mRMR-SVM 不仅在公共数据集 CCLE、GDSC, 以及三类特定组织中取得了较好的分类预测结果, 而且能够识别与抗癌药物反应相关联的生物标志物, 说明其可以作为抗癌药物反应分类预测的有效工具。此外, mRMR-SVM 具有可拓展性, 可融入其他类型数据 (如基因突变等) 进一步提升模型的预测性能。对于模型筛选的特征基因, 可以构建特征基因信息与药物敏感性之间的回归模型 (如岭回归、逻辑回归), 通过回归系数挖掘抗癌药物敏感性预测因子。

附件 PIBB\_20210082-prgm-S1.zip 请见本文网络版 ([www.pibb.ac.cn](http://www.pibb.ac.cn) 或 [www.cnki.net](http://www.cnki.net))。

## 参 考 文 献

- [1] Reardon S. Precision-medicine plan raises hopes. *Nature*, 2015, **517**(7536): 540-540
- [2] Dexter D L, Leith J T. Tumor heterogeneity and drug resistance. *Clin Oncol*, 1986, **4**(2): 244-257
- [3] Workman P, Clarke P A, Allazikani B. Personalized medicine: patient-predictive panel power. *Cancer Cell*, 2012, **21**(4): 455-458
- [4] Garraway L A. Genomics-driven oncology: framework for an emerging paradigm. *Clin Oncol*, 2013, **31**(15): 1806-1814
- [5] Caponigro G, Sellers W R. Advances in the preclinical testing of cancer therapeutic hypotheses. *Nat Rev Drug Discov*, 2011, **10**(3): 179-187
- [6] Garnett M J, Edelman E J, Heidorn S J, *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 2012, **483**(7391): 570-575
- [7] Barretina J, Caponigro G, Stransky N, *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 2012, **483**(7391): 603-307
- [8] Zhang N, Wang H, Fang Y, *et al.* Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol*, 2015, **11**(9): e1004498
- [9] Liu C, Wei D, Xiang J, *et al.* An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol Ther Nucleic Acids*, 2020, **21**: 676-686
- [10] Zhang L, Chen X, Guan N N, *et al.* A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction. *Front Pharmacol*, 2018, **9**: 1071
- [11] Xu X, Gu H, Wang Y, *et al.* Autoencoder based feature selection method for classification of anticancer drug response. *Front Genet*, 2019, **10**: 233
- [12] Hejase H A, Chan C. Improving drug sensitivity prediction using different types of data. *CPT Pharmacometrics Syst Pharmacol*, 2015, **4**(2): 98-105
- [13] Wang Y, Fang J, Chen S. Inferences of drug responses in cancer cells from cancer genomic features and compound chemical and therapeutic properties. *Sci Rep*, 2016, **6**(3): 32679
- [14] Su R, Liu X, Wei L, *et al.* Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods*, 2019, **166**: 91-102
- [15] 李军怀, 付静飞, 蒋文杰, 等. 基于 mRMR 的文本分类特征选择方法. *计算机科学*, 2016, **43**(1): 225-228
- [15] Li J H, Fu J F, Jiang W J, *et al.* *Computer Science*, 2016, **43**(1): 225-228
- [16] Wei D, Liu L, Zheng X, *et al.* Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics*, 2019, **20**(1): 44
- [17] 王嘉新, 胡增涛, 关沧海, 等. TP73-AS1 在肿瘤中的表达及调控作用. *中国病理生理杂志*, 2019, **35**(4): 764-768
- [17] Wang J X, Hu Z T, Guan C H, *et al.* *Chin J Pathophysiol*, 2019, **35**(4): 764-768
- [18] Jin W. Novel insights into PARK7 (DJ-1), a potential anti-cancer therapeutic target, and implications for cancer progression. *J Clin Med*, 2020, **9**(5): 1256
- [19] Yun J J, Kyeong L, Hyo J L, *et al.* Cationic amino acid transporter PQLC 2 is a potential therapeutic target in gastric cancer. *Cancer Sci*, 2019, **110**(4): 1453-1463
- [20] 古婉仪, 杨毅. IFI6 的研究现状. *临床与病理杂志*, 2015, **35**(10): 1847-1851
- [20] Gu W Y, Yang Y. *J Clin Pathol Res*, 2015, **35**(10): 1847-1851

## A Computational Model for Predicting Classification of Anticancer Drug Response to Individual Tumor and Its Applications\*

LI Shao-Da, LI Yu-Shuang\*\*

(School of Science, Yanshan University, Qinhuangdao 066004, China)

**Abstract Objective** Different patients may have different responses to the same anticancer drug. Understanding differences in anticancer drug responses among patients is crucial for cancer precision medicine. **Methods** High-throughput sequencing data make it possible to construct anti-cancer drug response classification models. Based on two classic data sets, Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC), this paper proposed a classification model, mRMR-SVM, by employing Max-Relevance and Min-Redundancy (mRMR) algorithm and Support Vector Machine (SVM). Feature genes were extracted by using variance ranking and mRMR algorithm on gene expression data, and SVM was applied to predict that an anticancer drug is sensitive or resistant to a given cell line. **Results** The experimental results showed that the average accuracy of mRMR-SVM is 0.904 for 22 drugs in CCLE, and 0.851 for 11 drugs in GDSC, higher than traditional SVM, Random Forest, Deep Response Forest, Deep Neural Network and CDCN. **Conclusion** mRMR-SVM also has good generalization due to its satisfactory classification prediction on three specific tissues. In addition, mRMR-SVM could identify biomarkers closely related to the occurrence and development of cancer.

**Key words** anticancer drug response, max-relevance and min-redundancy, support vector machine, sensitive, resistant

**DOI:** 10.16476/j.pibb.2021.0082

---

\* This work was supported by grants from Natural Science Foundation of Hebei, China (A2020203021) and Science Foundation for Returned Scholars of Hebei Province (C20200365).

\*\* Corresponding author.

Tel: 86-15232356135, E-mail: yushuangli@ysu.edu.cn

Received: April 1, 2021 Accepted: May 10, 2021