·新技术与新方法·

基于AdaBoost算法的药物—靶向蛋白作用预测算法

古万荣[1,2]，谢贤芬[2,3]，何亦琛[1]，张子烨[4]

1. 华南农业大学 数学与信息学院 计算机科学与工程系（广州 510642）
2. 广东省大数据分析与处理重点实验室（广州 510006）
3. 暨南大学 经济学院 统计系（广州 510632）
4. 华南理工大学 数学学院 数学系（广州 510660）

【摘要】 对靶向蛋白的药物作用进行预测可以促进药物新作用的发现。新近的研究更倾向于单独将特定的矩阵填补算法应用在靶向蛋白和药物的相互作用预测中。单模型的矩阵填补算法准确度较低，因此应用在药物—靶向蛋白作用预测方面也难以获得满意的结果。AdaBoost 算法是一种由多分类器组合生成强分类器的算法框架，其在分类应用领域的实用性和有效性已被证明。靶向蛋白的药物作用预测是一个矩阵填补问题，即是一种评分预测过程，因此本文在使用 AdaBoost 算法对药物—靶向蛋白作用进行预测前，将药物—靶向蛋白作用预测的矩阵填补问题转化为分类问题，将 AdaBoost 算法应用在靶向蛋白的药物作用预测评分上，充分利用 AdaBoost 算法框架对多个弱分类器进行融合从而提升性能，得以进行准确的药物—靶向蛋白作用预测。基于公测数据集的实验结果表明，本文提出的算法在预测准确度方面超过了大多数经典算法和新近算法，较好地克服了新近基于机器学习方法单算法的局限性，更好地挖掘隐含因素，有效提升了预测准确度。

【关键词】 靶向蛋白；药物作用预测；评分预测；AdaBoost 算法

Drug-target protein interaction prediction based on AdaBoost algorithm

GU Wanrong[1,2], XIE Xianfen[2,3], HE Yichen[1], ZHANG Ziye[4]

1. Department of Computer Science and Engineering, School of mathematics and information, South China Agricultural University, Guangzhou 510642, P.R.China
2. Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou 510006, P.R.China
3. Department of Statistics, School of Economy, Jinan University, Guangzhou 510632, P.R.China
4. Department of Mathematics, School of Mathematics, South China University of Technology, Guangzhou 510660, P.R.China
Corresponding author: XIE Xianfen, Email: txiexianfen2009@jnu.edu.cn

【Abstract】 The drug-target protein interaction prediction can be used for the discovery of new drug effects. Recent studies often focus on the prediction of an independent matrix filling algorithm, which apply a single algorithm to predict the drug-target protein interaction. The single-model matrix-filling algorithms have low accuracy, so it is difficult to obtain satisfactory results in the prediction of drug-target protein interaction. AdaBoost algorithm is a strong multiple classifier combination framework, which is proved by the past researches in classification applications. The drug-target interaction prediction is a matrix filling problem. Therefore, we need to adjust the matrix filling problem to a classification problem before predicting the interaction among drug-target protein. We make full use of the AdaBoost algorithm framework to integrate several weak classifiers to improve performance and make accurate prediction of drug-target protein interaction. Experimental results based on the metric datasets show that our algorithm outperforms the other state-of-the-art approaches and classical methods in accuracy. Our algorithm can overcome the limitations of the single algorithm based on machine learning method, exploit the hidden factors better and improve the accuracy of prediction effectively.

【Key words】 target protein; drug effect prediction; score prediction; AdaBoost algorithm

DOI：10.7507/1001-5515.201802026
基金项目：国家重点研发计划重点专项（2017YFC1601701）；广东省大数据分析与处理重点实验室开放基金项目（2017006，2017010）
通信作者：谢贤芬，Email：txiexianfen2009@jnu.edu.cn

引言

在药物发现研究中，对药物—靶向蛋白相互作用（drug-target interaction，DTIs）预测是一个重要的研究热点。当前，一种新药物的发现仍然是一个非常昂贵且低效的过程，大约每一个新分子实体（new molecular entity，NME）需要花费 18 亿美元的研发成本[1]。此外，新药从研发到市场销售往往需要花费约 10 年时间。举个例子，每年只有约 20 种新药被美国食品和药品监督管理局（food and drug administration，FDA）批准。近年来，成功获批的新药数量在逐年下降。在上述情况下，对旧药物的新作用的挖掘成为一个新的研究方向。

随着药物—靶向蛋白作用研究的深入，相关的公测数据库和研究在逐年增加。例如，DrugBank 数据库具有丰富的生物和化学标注数据[2]，且具有详细的药物数据资料，包括药物类型、药物简介、化学结构、药物成分、药物靶点、药物相互作用等，该数据库由加拿大卫生研究院提供支持。此外，化学相互作用搜索工具（search tool for interactions of chemicals，STITCH）是一种预测化学—蛋白相互作用的公测数据集[3]。STITCH 数据库的关联数据主要来源于高通量实验、基因组上下文预测、共同表达、自动化文本挖掘等，更详细的信息参见本文实验部分。新近的研究表明，经 FDA 批准的药物中约 96% 标注有靶向作用蛋白。然而，仍然有许多靶向蛋白和其他药物的作用关系是未知的。

过去数十年，有许多学者对药物—靶向蛋白进行了系统研究，而传统的药物发现主要遵循“一个分子，一个目标，一种疾病”的范式，这种历史范式确定了一些影响特定蛋白质的有效化学分子。然而，这种传统的药物发现范式的主要局限性在于，药品的设计目标是针对疾病系统中的个体因素，但是复杂的疾病在本质上是多因素的，即疾病及症状是不同基因和途径的突变和干预的逐渐积累，在这种多重因素中药物发现的准确性很容易受到影响。传统模式忽视了药物与其靶向蛋白之间复杂的相互作用，而许多复杂疾病往往与多种目标蛋白有关，因此这种模式并没有像预期的那样加速药物发现的速度[4-5]。大多数药物靶点是细胞蛋白，其目的是通过与化合物有选择性的相互作用来治疗或诊断疾病。目前的研究表明，经典的治疗性药物靶标包含 130 个蛋白质家族，如酶、G 蛋白偶联受体（G-protein-coupled receptors，GPCRs）、离子通道和转运体、核激素受体等[6-7]。据估计，人类基因组中约有 6 000 ~ 8 000 个具有药理学价值的靶点，但是到目前为止，这些目标中只有小部分验证了与已批准的药物之间的相互作用，大量假定的药物靶点仍有待验证[7-9]。由于药物和蛋白的特征的高维性和复杂性，新近研究提出使用机器学习方法对药物和蛋白的相互作用进行自动挖掘，借助于数值分析和挖掘的手段，将药物和靶向蛋白作用问题转化为协同过滤、矩阵分解或评分预测问题，提高了药物和靶向蛋白作用预测的效率。虽然相对于传统药物发现，使用机器学习方法已经具备明显优势，且随着人工智能和高性能计算的发展，借助于计算机手段进行该领域的研究和应用也将成为未来的发展方向，但就目前而言，该领域内的研究仍然存在至少以下两个问题：

（1）在单个机器学习算法中，很难将复杂的参数和隐含的作用关联挖掘出来[10]；另外，在单个机器学习算法中，如果模型过于简单，则会影响预测准确率，如果过于复杂，即参数过多，则会产生过拟合问题。

（2）药物冷关联问题仍然无法在基于矩阵填补思想的预测算法中得到很好的解决[11]；冷关联问题，是指在推荐算法或评分预测研究中，存在部分对象和其他对象关联很少的现象，如果冷关联太多，对预测算法具有较大影响。

为了解决以上问题，本文提出一种较高效的基于 AdaBoost 算法的提升方法，即将多个弱分类器组合在一起，形成一个强的分类器。同时，本文通过对 AdaBoost 算法进行改进，将评分预测问题转化为分类问题，将算法框架应用到药物—靶向蛋白相互作用预测的矩阵填补研究中。本文提出的方法利用机器学习算法进行药物—靶向蛋白作用预测，相较于传统方法，具有高效、低成本的优势，同时也较好地克服了新近基于机器学习方法的单算法的局限性，能够更好地挖掘隐含因素，有效提升预测准确度，在药物冷关联问题方面，也比新近方法更好。

1　相关研究介绍

新近的药物—靶向蛋白作用预测可以大致分为 3 类：基于网络模型的方法、基于机器学习的方法以及其他模型方法。

（1）基于网络模型的方法

基于网络模型的方法是新近比较常用的药物—靶向蛋白作用预测方法。周福家等[12]提出了寻找多目标优化解决方法（multiple target optimal interven-

tion，MTOI）来获得最佳的疾病转换流程，通过最佳的疾病转换流程的挖掘来确定药物—靶向蛋白作用。Campillos 等[11]提出了一种基于药物边缘效应相似度的方法来确定两种药物是否对于同一种靶向蛋白具有相似作用。Chen 等[4]提出基于网络模型的随机游走方法来挖掘药物—靶向蛋白作用。

（2）基于机器学习的方法

最近，有研究将药物—靶向蛋白的点对标注为“正”、“负”标签样本，然后进行有监督的学习预测。这种情况下，“负”标签没有作用关联，因此不能确定“负”标签的标注。另外一种常见的基于机器学习的模型是将药物—靶向蛋白看作二分图模型，然后基于二分图进行挖掘，Yamanishi 等[13]提出了一种核回归的方法来挖掘药物—靶向蛋白的二分图。

（3）其他模型方法

其他模型方法包括：Keiser 等[14]提出了使用化学相似度来预估药物—靶向蛋白的相互作用关系；另外一种常用方法是使用基于概率矩阵分解的方法（probabilistic matrix factorization，PMF）来预测药物—靶向蛋白关系[15]。

2 理论基础原理

2.1 AdaBoost 算法

AdaBoost 算法是一个经典的自适应集成算法，其基本原理在于通过对弱可学习的算法进行集成，进而生成强可学习的模型。该算法需要有输入数据集 $\boldsymbol{D}$，若干个基础分类模型集合 $\boldsymbol{C}$。该算法首先通过对若干个训练样本的学习训练获得第一个弱分类器；然后，将错误分类的样本和其他未训练样本一起构成新的训练样本集，通过再次训练获得第二个弱分类器；再然后，将以上两次学习都分类错误的样本加上未学习的样本一起构成新的训练样本，进而得到第三个弱分类器；以此类推，通过若干次迭代后，该分类器得到性能加强和提升，逐渐变成强分类器[10]。现在，AdaBoost 算法已被 Schapire[16]证明弱可学习和强可学习是等价的。

2.2 矩阵分解模型

矩阵分解模型起源于代数中的奇异值分解，是挖掘具有二分图结构的推荐问题的重要工具之一，它通过矩阵降维的方式将原本较稀疏的矩阵缺失数据进行补全。在挖掘二分图结构关系中，可以将药物—靶向蛋白的作用关系矩阵分解为两个较低维度的矩阵的乘积，如式（1）所示：

$$\hat{\boldsymbol{R}}=\mathbf{Q}^{\mathrm{T}}\boldsymbol{P} \tag{1}$$

其中，$\boldsymbol{P}\in\mathbb{R}^{k\times m}$，$\mathbf{Q}\in\mathbb{R}^{k\times n}$ 是两个降维后的矩阵。$\boldsymbol{P}$ 矩阵包含了全部药物及其隐含属性，$\boldsymbol{Q}$ 矩阵包含了全部靶向蛋白及其隐含属性。k 是隐含因子的维度，大小可根据经验设定。其物理意义是：假设每个药物都有 k 个隐含的属性，这些属性可以和靶向蛋白的 k 隐含属性相对应，使之产生作用关联。因此，药物对靶向蛋白的作用可以用 $\hat{\boldsymbol{R}}(d,t)=\hat{r}_{dt}$ 表示，如式（2）所示：

$$\hat{r}_{dt}\approx\sum_{k=1}^{n}p_{dk}q_{tk} \tag{2}$$

其中，$p_{dk}=\boldsymbol{P}(d,k)$，$q_{tk}=\boldsymbol{Q}(t,k)$ 在现实场景中，可能会面临更复杂的情况，因此应该对模型进行适当的修正，加入偏差系数等。

3 基于 AdaBoost 算法的药物—靶向蛋白作用预测算法

在推荐系统的评分预测模型中，一般先将用户集合和物品集合分别定义为：$\boldsymbol{U}=\{u_1,u_2,\cdots,u_{|U|}\}$，$\boldsymbol{I}=\{i_1,i_2,\cdots,i_{|I|}\}$。并以此为基础定义评分矩阵，即：$\boldsymbol{D}=\{(u,i,r_{ui})_1,(u,i,r_{ui})_2,\cdots,(u,i,r_{ui})_m\}$，共计 m 个评分，一般而言，有效的评分数 $m\ll|\boldsymbol{U}|\times|\boldsymbol{I}|$。

然而，传统的 AdaBoost 算法框架适用于分类问题，如前文的算法框架伪代码。如要将其应用到药物—靶向蛋白作用预测中，则需要将矩阵分解模型转化为分类模型，因此需要对 AdaBoost 算法模型进行改进，分别需要解决以下问题：

（1）如何将评分预测问题转化为分类问题，即将药物—靶向蛋白作用预测的二值矩阵分解和填补问题转化为 AdaBoost 算法框架可以应用的问题。

（2）模型错误率的评估和计算。

（3）训练数据的分类权重分布如何与矩阵分解模型结合。

（4）如何将多个模型集成起来。

3.1 问题转化

AdaBoost 算法对多弱分类模型集成的性能已经得到较多研究成果确认。因此，本文将药物—靶向蛋白作用预测问题转化为二值分类问题。首先，引入一个参数 ϕ 作为阈值，该阈值的作用是用于评判矩阵分解模型中的评分预测结果是好、还是坏。假设某药物对某靶向蛋白的作用预测结果为 3，通过矩阵分解模型得出的分值如果在$[3-\phi,3+\phi]$区间内，则认为该矩阵分解模型是比较准确的，反之则认为该模型不准确。通过这种映射，将矩阵分解的评分预测模型转化为二值分类问题。

3.2 模型错误率评估

分类方法的模型错误率很容易通过错误分类样本进行统计，评分预测问题的错误率评估往往通过预测结果 $\hat{r}_{dt}$ 和真实结果 r_{dt} 的差距来表示，比如以均方根误差（root mean square error，RMSE）或平均绝对误差（mean absolute error，MAE）等来表示。在本文提出的方法处理中，当预测结果 $\hat{r}_{dt} \in [r_{dt}-\phi, r_{dt}+\phi]$ 时，被认为是准确的，即不计算其错误率，否则使用以下计算，如式（3）所示：

$$\varepsilon_i = \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt} \tag{3}$$

公式中，$\boldsymbol{d}$ 表示药物集合，$\boldsymbol{t}$ 表示靶向蛋白集合。

3.3 权重更新

AdaBoost 算法的有效性很大程度上依赖于各种基础模型的权重分布，即对错误的分类数据会加大权重，在下一轮迭代中会更加重视其训练，反之则降低权重，不断迭代后，分类器逐渐更具鲁棒性。最后将所有模型集成起来以达到最优效果。因此，本文定义损失函数如式（4）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} (r_{dt}-\hat{r}_{dt})^2 \tag{4}$$

将损失函数加入到每个训练数据中以增强训练效果，如式（5）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt}(r_{dt}-\hat{r}_{dt})^2 \tag{5}$$

本文使用随机梯度下降法来实现药物—靶向蛋白作用矩阵中的预测问题，即将权重赋值给如式（6）所示：

$$\arg\min \sum_{d=1}^{|\boldsymbol{d}|}\sum_{t=1}^{|\boldsymbol{t}|} w_{dt}[(r_{dt}-\mu-\boldsymbol{b}_t-\boldsymbol{b}_d-\boldsymbol{q}_t{}^{\mathrm{T}}\boldsymbol{p}_u)^2 + \lambda(\boldsymbol{b}_t{}^2+\boldsymbol{b}_d{}^2+||\boldsymbol{q}_i||^2+||\boldsymbol{p}_u||^2)] \tag{6}$$

其中，w_{dt} 是药物 d 对于靶向蛋白 t 的作用评分的权重。每个因子的递推迭代公式如式（7）所示：

$$\begin{aligned} \boldsymbol{b}_t &\leftarrow \boldsymbol{b}_t + w_{dt}\cdot\gamma\cdot(e_{dt}-\lambda\cdot\boldsymbol{b}_t) \\ \boldsymbol{b}_d &\leftarrow \boldsymbol{b}_d + w_{dt}\cdot\gamma\cdot(e_{dt}-\lambda\cdot\boldsymbol{b}_d) \\ \boldsymbol{q}_t &\leftarrow \boldsymbol{q}_t + w_{dt}\cdot\gamma\cdot(e_{dt}\boldsymbol{p}_u-\lambda\cdot\boldsymbol{q}_t) \\ \boldsymbol{p}_u &\leftarrow \boldsymbol{p}_u + w_{dt}\cdot\gamma\cdot(e_{dt}\boldsymbol{q}_t-\lambda\cdot\boldsymbol{p}_u) \end{aligned} \tag{7}$$

根据前面的错误率，权重更新公式如式（8）所示：

$$\alpha_i = \mu\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right) \tag{8}$$

其中，参数 μ 是调整参数，用于自适应训练数据的调整。

3.4 模型集成

通过 AdaBoost 算法展开训练，可以集成 T 个基础训练模型的融合模型，并通过加权平均的方法来集成各个模型的药物—靶向蛋白作用预测框架，每个模型有一个对应权重 α_i，最终的药物—靶向蛋白作用预测结果如式（9）所示：

$$\hat{\boldsymbol{R}}_{dt}(d,t) = \sum_{i=1}^{T}\alpha_i h_i(d,t) \tag{9}$$

3.5 算法设计与分析

基于前文讨论，AdaBoost 算法框架的药物—靶向蛋白作用预测算法具体如下：

输入：数据集 $\boldsymbol{D}=\{(d,t,r_{dt},w_{dt})_1,(d,t,r_{dt},w_{dt})_2,\cdots,(d,t,r_{dt},w_{dt})_m\}$；基础评分预测模型 L，如奇异值分解方法（singular value decomposition，SVD）或概率矩阵分解（probabilistic matrix factorization，PMF）等；训练轮数 T；更新权重的控制参数 μ；

输出：$\hat{r}_{dt}(d,t) = \sum_{i=1}^{T}\alpha_i h_i(d,t)$

1：$(d,t,r_{dt},w_{dt})_t^1 = \frac{1}{m}$；

2：**for** $i=1,\cdots,T$ **do**

3：$h_i = L(\boldsymbol{D},\boldsymbol{D}_i)$；

4：**for** all $\hat{r}_{dt} \notin [r_{dt}-\phi, r_{dt}+\phi]$ in $\boldsymbol{D}_i$, the set is $\boldsymbol{D}_i^-$ **do**

5：$\varepsilon_i = \sum_{d,t\in D_i^-} w_{dt}^t$；

6：**end for**

7：$\alpha_i = \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$；

8：$\boldsymbol{D}_{i+1}(t) = \frac{\boldsymbol{D}_i(t)}{Z_i}\times\begin{cases}\exp(\alpha_i)\ if\ h_i(x_t)\neq y_t \\ \exp(-\alpha_i)\ if\ h_i(x_t)=y_t\end{cases}$；

9：**end for**

上述算法有明确的输入，即需要基本的算法模型，如 SVD 奇异值分解算法；其次，需要有训练集 $\boldsymbol{D}$、训练轮数 T、判断作用分准确与否的阈值 ϕ 以及更新权重的控制参数 μ。

4 实验与结果分析

本文实验基于两个公测数据集：人工标注靶点和药物在线资源（manually annotated targets and drugs online resource，MATADOR）和前文“引言”部分提及的 STITCH 数据集[3, 17]，对比分析了方法中的参数及对预测结果的影响，并针对若干经典分类方法和新近方法进行仿真分析。

4.1 数据集与评测指标

本文的实验数据集主要来自两个公开数据集：① MATADOR（网址为：http://matador.embl.de/）[17]，是一个公开的来源于人工整理和自动文本挖掘的在线治疗数据库，其中包含了许多药物—靶向蛋白的相互作用关系数据。本文实验选用了该数据库中具有药物—靶向蛋白直接或间接作用关系的数据。MATADOR 数据库中包含了 784 种药物，2 431 种靶向蛋白和 13 064 个药物—靶向蛋白相互作用，其中共包含了 7 862 种直接和 5 202 种间接作用关系。② STITCH（网址为：http://stitch.embl.de/）[3]，该数据库提供了许多靶向蛋白及其化学成分的相互关系数据，这些标注关系来源于已知的实验和文献，具有较高的可信度。本文选用了 STITCH 数据库中和 MATADOR 数据库重合的药物—靶向蛋白数据。STITCH 数据库中包含了 598 种药物，671 种靶向蛋白和 3 296 个有效的药物—靶向蛋白作用对，在其中分别有 2 589、945 和 1 493 条相互作用关系，分别对应于绑定、激活和隐含关系。为了能在作用预测中评分更直观，本文将数据集中的作用分数均匀映射到[0, 5]之间。

本文实验将上述数据集平均分为 5 份，然后使用 5 折交叉验证方法进行实验，其中 1 份作为测试集，剩余 4 份作为训练集。由于药物—靶向蛋白作用预测是一种评分预测问题，因此常使用 RMSE（以符号 *RMSE* 表示）方法作为评价指标[18]，定义为预测值与真实值之间的差的平方均值，如式（10）所示：

$$RMSE = \frac{\sqrt{\sum_{d=1}^{|d|}\sum_{t=1}^{|t|}(r_{dt}-\hat{r}_{dt})^2}}{|\boldsymbol{D}|} \tag{10}$$

与 RMSE 评测标准同样常用的是 MAE，区别在于 RMSE 对于错误作用预测的惩罚扣分会更多，即对错误结果更苛刻。

本文还使用了预测准确率（accuracy，Accu）和召回率（recall，Rec）两个指标。这两个指标的计算依赖于实际样本数据和预测统计数据，如表 1 所示，真阳性（true positive，TP）（以符号 *TP* 表示），指具体的药物—靶向蛋白具备相互作用且预测结果显示也具备相互作用的统计结果总数。以此类推，假阴性（false negative，FN）（以符号 *FN* 表示），指药物—靶向蛋白具备相互作用，但预测结果显示无相互作用的统计结果总数；假阳性（false positive，FP）（以符号 *FP* 表示），指药物—靶向蛋白不具备相互作用，但预测结果显示有相互作用的统计结果总数；真阴性（true negative，TN）（以符号 *TN* 表示），指药物—靶向蛋白不具备相互作用，预测结果也显示无相互作用的统计结果总数。

通过以上的统计数据，Accu（以符号 *Accu* 表示）的计算结果如式（11）所示：

$$Accu = (TP+TN)/(TP+TN+FN+FP) \tag{11}$$

Rec（以符号 *Rec* 表示）的计算结果如式（12）所示：

$$Rec = TP/(TP+FN) \tag{12}$$

4.2 实验结果与分析

矩阵分解模型是评分预测模型中常用的方法，因此，本文的 AdaBoost 算法框架使用 SVD 方法和 PMF 模型作为药物—靶向蛋白作用预测的基本预测模型。矩阵分解过程中的学习速率和惩罚系数是重要的预设参数，根据模型训练经验，本文在 SVD 方法中设定的学习速率为 0.01、惩罚系数为 0.01，对于 PMF 模型，设定学习速率为 0.005、惩罚系数为 0.02。

4.2.1 AdaBoost 算法框架对比分析　如表 2 所示，展示了本文提出的 AdaBoost 算法在药物—靶向蛋白中的作用预测方法。从实验结果可知，使用了 AdaBoost 算法相对于单个模型而言，效果有了较大的提升。

从实验结果可知，在两个不同的数据集和不同的基础模型中，使用了 AdaBoost 算法的运算结果的误差较小。其原因在于，使用 AdaBoost 算法将基础方法进行融合提升，通过对分类错误进行二次迭代后不断逼近，以此达到预测准确度提升的预期效果。

4.2.2 算法框架的相关参数分析　本文所提方法的参数主要有阈值 ϕ，训练轮数 T，权重更新 μ，本实

表 1　药物—靶向蛋白作用预测的统计结果

Tab.1　Statistical results of drug-target protein interaction prediction

	预测有作用	预测无作用
实际有作用	*TP*	*FN*
实际无作用	*FP*	*TN*

表 2　AdaBoost 算法实验效果

Tab.2　Experimental results of AdaBoost algorithm

算法	RMSE			
	MATADOR		STITCH	
	未使用	AdaBoost	未使用	AdaBoost
SVD	0.923 4	**0.892 2**	0.832 1	**0.807 6**
PMF	0.943 8	**0.912 3**	0.859 8	**0.832 1**

验基于 STITCH 数据集，使用 SVD 方法进行实验，实验结果如图 1 所示。

从以上实验结果可知，阈值 ϕ 的选取对 RMSE 实验结果的影响是先大再小，然后再变大，阈值 ϕ 在[0.1，1.0]取值的 10 个实验数据中，当取 0.7 时，使得 RMSE 值最小。所以，如果想要得到最好的预测准确度，应当调整适当的阈值 ϕ，过大或过小都会对预测造成不良影响。

如图 2 所示，随着训练轮数的增长，一开始精度提高较快，达到 7 次以上时，精度略有提高，但趋势变缓。

如图 3 所示，权重更新参数 μ 对药物—靶向蛋白作用预测也具有明显的影响，从实验可知，μ = 0.4 时，准确度最高。

4.2.3 *其他方法的准确度对比与分析*　本实验实现了一些经典的评分预测方法和新近方法作为基线方法进行对比，具体如下：

基于近邻用户的推荐算法（user-based）[19]：该算法应用到药物—靶向蛋白预测中，将近邻药物看作是近邻用户。

基于近邻物品的推荐算法（item-based）[20]：该算法应用到药物—靶向蛋白预测中，将近邻蛋白看作是近邻物品。

朴素贝叶斯算法（Naïve Bayes）[21]：是一种基于贝叶斯后验概率的分类预测方法，常用于数据分类和评分预测。

支持向量机（support vector machine，SVM）方法[22]：这是一种性能比较优越的分类器模型，该算法使用超平面进行分类训练，对非线性数据也可以采用核技术将其进行准确分类[23]。

MTOI（周福家等[12]提出）：通过寻找多目标优化解决方法来获得最优的疾病转换状态，通过最优的疾病转换状态的挖掘来确定药物—靶向蛋白作用。

基于二分图的核回归方法（Yamanishi 等[13]提出）：该方法通过将药物—靶向蛋白关系映射为二分图，然后通过核回归方法挖掘二分图的待预测节点。

基于化学相似度的方法（Keiser 等[14]提出）：该方法通过蛋白质的化学相似度作为挖掘基础，从而推断药物—靶向蛋白的作用关系。

以上方法的最终实验结果如表 3 所示，“本文方法”是指本文使用 AdaBoost 算法的结果，基础模型使用了 SVD 方法。

从表3实验结果可知，本文算法获得了较好结果，其原因在于，本文使用 AdaBoost 算法将基础方法如 SVD 方法进行迭代融合，多轮迭代更逼近准确结果。从表 2 也可以看出，即便 SVD 方法不使用模型融合也获得不差的结果，在使用了 AdaBoost 算法后，准确度明显上升。

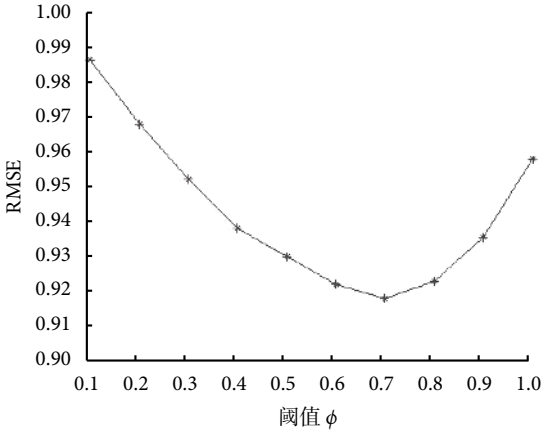

图 1　阈值 ϕ 对预测结果的影响

Fig.1　The influence of threshold ϕ

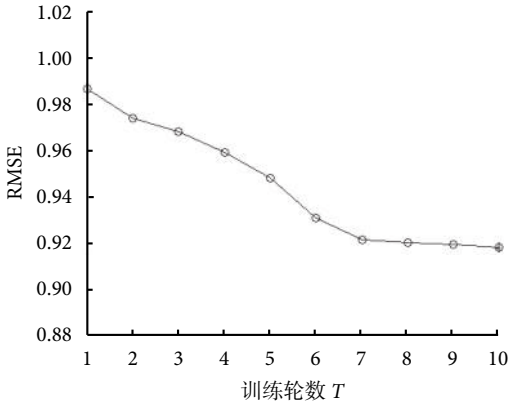

图 2　训练轮数 T 对预测结果的影响

Fig.2　The influence of training iterations T

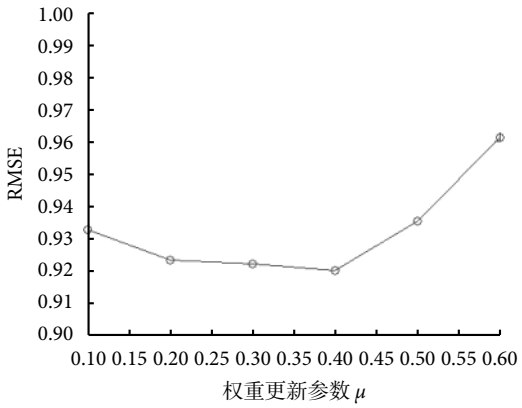

图 3　权重更新参数 μ 对预测结果的影响

Fig.3　The influence of weight updating parameter μ

4.2.4 *冷关联问题的实验*　在药物—靶向蛋白作用预测中，本文发现在训练集中，有些药物关联的蛋

表 3 本文方法与新近方法的实验结果对比

Tab.3 Comparison of recent approaches and our proposed method

算法	数据集	Accu	Rec	RMSE
user-based 方法	MATADOR	89.12%	85.12%	1.115
	STITCH	88.54%	84.62%	1.295
item-based 方法	MATADOR	88.45%	84.12%	1.138
	STITCH	86.41%	82.89%	1.345
Naïve Bayes 方法	MATADOR	91.52%	88.45%	1.038
	STITCH	89.92%	86.95%	1.120
SVM 方法	MATADOR	91.92%	89.21%	0.914
	STITCH	90.10%	88.32%	0.983
MTOI 方法	MATADOR	91.35%	90.35%	0.902
	STITCH	90.65%	89.65%	0.972
基于二分图的核回归方法	MATADOR	92.36%	91.62%	0.926
	STITCH	91.69%	90.62%	0.982
基于化学相似度的方法	MATADOR	90.18%	89.12%	0.933
	STITCH	88.64%	85.64%	0.950
本文方法	MATADOR	**95.10%**	**93.21%**	**0.807**
	STITCH	**94.58%**	**91.58%**	**0.892**

表 4 本文方法与新近方法在冷关联数据集上的实验结果对比

Tab.4 Comparison of recent approaches and our proposed method in cold relevance data set

算法	数据集	Accu	Rec	RMSE
user-based 方法	MATADOR-C	23.38%	20.56%	2.358
	STITCH-C	22.78%	19.78%	2.669
item-based 方法	MATADOR-C	22.89%	21.65%	2.610
	STITCH-C	20.36%	19.35%	2.721
Naïve Bayes 方法	MATADOR-C	25.45%	24.36%	2.221
	STITCH-C	19.96%	20.39%	2.389
SVM 方法	MATADOR-C	25.92%	26.75%	2.003
	STITCH-C	20.17%	**25.96%**	2.169
MTOI 方法	MATADOR-C	25.65%	24.98%	1.985
	STITCH-C	24.89%	23.54%	2.069
基于二分图的核回归方法	MATADOR-C	27.11%	26.52%	1.845
	STITCH-C	26.05%	24.85%	1.896
基于化学相似度的方法	MATADOR-C	26.25%	25.59%	1.956
	STITCH-C	24.47%	22.84%	2.056
本文方法	MATADOR-C	**29.02%**	**26.87%**	**1.782**
	STITCH-C	**27.54%**	24.87%	**1.806**

白非常少，这就是评分预测研究中常见的“冷关联问题”，冷关联问题加剧了数据挖掘中的数据稀疏性的负面影响。本文对实验选取的数据集进一步整理，形成关联较少的数据集，以验证本文方法对冷关联数据集下的有效性。首先，对 MATADOR 数据集和 STITCH 数据集进行整理，挑选了药物关联数少于 3 的药物、蛋白以及关联信息。经过挑选后，MATADOR 数据库中包含了 365 种药物，958 种靶向蛋白和 5 211 个有效的药物—靶向蛋白作用对，其中共包含了 3 851 种直接和 1 860 种间接作用关系。STITCH 数据库中包含了 231 种药物，158 种靶向蛋白和 1 213 个有效的药物—靶向蛋白作用对，在其中分别有 512、254 和 447 条相互作用关系，分别对应于绑定、激活和隐含关系。为了便于区分，挑选后的数据库分别称为 MATADOR—冷关联（cold）（以符号 MATADOR-C 表示）和 STITCH—冷关联（cold）（以符号 STITCH-C 表示）。实验结果如表 4 所示。

从实验结果可知，冷关联数据集中的实验结果普遍都不好，不管是 Accu、Rec 还是 RMSE，各种方法都受到很大影响。本文方法除了在 STITCH-C 数据集上的 Rec 指标中，比 SVM 方法较差外，其余实验结果均获得较好结果，但该指标中的结果数据相差也较微小。

5 结语

本文提出使用融合学习模型的 AdaBoost 算法来提高药物—靶向蛋白作用预测的效果，首次在该领域使用这种融合算法。实验结果表明，本文方法具有较高的预测准确度。使用 AdaBoost 算法可以让弱分类器融合组成较强分类方法，提升分类性能，使用本文方法还能将作用预测矩阵切换为分类算法框架。本文工作是对目前新近算法常使用单机器学习模型进行药物—靶向蛋白作用预测的一种多模型融合尝试。将来的研究工作中，可能还会面临两个方面的问题函待解决：① 隐含关联特征如何进行有效提取及其如何确定其在相互作用挖掘中的权重。对于数据挖掘中进行特征工程研究，业界已经有了深度学习方法（deep learning，DL）及其相关的软硬件平台体系，并在人工智能等领域取得了良好的应用效果。如何将 DL 应用在药物—靶向蛋白作用预测将是未来的一个重要研究方向。② 海量数据的快速处理。随着药物和病种亚型的增加，相互作用关联关系迅速增长，如何处理海量的关联关系也将是今后的重要研究方向。映射—降解（map-reduce，MR）处理框架是目前流行的海量数据批处理框架，在未来工作中，为更好地处理海量的药物—靶向蛋白、蛋白—蛋白等作用预测，本研究还将持续研究本文算法在 MR 框架中的分布式集群处理方法，使其可快速获得分析结果，并推进该研究的应用实践。在特征挖掘的研究方面，本研究将使用 DL 方法，深入挖掘隐含特征集的关系，以提升药物—靶向蛋白作用的预测准确度。

参考文献

1 Paul S M, Mytelka D S, Dunwiddie C T, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, 2010, 9(3): 203-214.

2 Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new

light on drug metabolism. Nucleic Acids Research, 2014, 42(Database issue): D1091-D1097.
3 Kuhn M, Szklarczyk D, Pletscher-Frankild S A, *et al*. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res, 2014, 42(D1): D401-D407.
4 Chen Xing, Liu Mingxi, Yan G Y. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst, 2012, 8(7): 1970-1978.
5 Iskar M, Zeller G, Zhao Xingming, *et al*. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Current opinion in biotechnology, 2012, 23(4): 609-616.
6 Hopkins A L, Groom C R. The druggable genome. Nature reviews Drug discovery, 2002, 1: 727-730.
7 Drews J. Drug discovery: A historical perspective. Science, 2000, 287(5460): 1960-1964.
8 Overington J P, Al-Lazikani B, Hopkins A L. How many drug targets are there? Nature Reviews Drug discovery, 2006, 5: 993-996.
9 Landry Y, Gies J P. Drugs and their molecular targets: an updated overview. Fundamental & clinical pharmacology, 2008, 22(1): 1-18.
10 Zhu Ji, Zou Hui, Rosset S, *et al*. Multi-class AdaBoost. Stat Interface, 2009, 2(3): 349-360.
11 Campillos M, Kuhn M, Gavin A C, *et al*. Drug target identification using side-effect similarity. Science, 2008, 321(5886): 263-266.
12 周福家, 张宏伟, 李卫国. 分子网络多靶标筛选的粒子群数值模拟法. 计算力学学报, 2015, 32(2): 269-273.
13 Yamanishi Y, Araki M, Gutteridge A A, *et al*. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics, 2008, 24(13): I232-I240.
14 Keiser M J, Setola V, Irwin J J, *et al*. Predicting new molecular targets for known drugs. Nature, 2009, 462(7270): 175-181.
15 Cobanoglu M C, Liu Chang, Hu Feizhuo, *et al*. Predicting drug-target interactions using probabilistic matrix factorization. J Chem Inf Model, 2013, 53(12): 3399-3409.
16 Schapire R E. The strength of weak learnability. Mach Learn, 1990, 5(2): 197-227.
17 Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics, 2009, 25(18): 2397-2403.
18 Takács D, Pilászy I, Németh B. Major components of the gravity recommendation system. ACM SIGKDD Explorations Newsletter, 2007, 9(2): 80-83.
19 Wu Zhenhua, Chen Xiaosu, Xiao Daoju. Offline Chinese signature verification based on segmentation and RBFNN classifier. Acta Automatica Sinica, 2007, 345(1): 995-1001.
20 Li Dongsheng, Chen Chao, Lv Qin, *et al*. An algorithm for efficient privacy-preserving item-based collaborative filtering. Future Generation Computer Systems, 2016, 55(C): 311-320.
21 Bilal M, Israr H, Shahid M, *et al*. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences, 2016, 28(3): 330-344.
22 Patel T B, Patil H A. Cochlear filter and instantaneous frequency based features for spoofed speech detection. IEEE J Sel Top Signal Process, 2017, 11(4): 618-631.
23 刘晓峰, 张雪英, Wang Z J. Logistic核函数及其在语音识别中的应用. 华南理工大学学报:自然科学版, 2015, 43(5): 100-106.

收稿日期：2018-02-10　修回日期：2018-07-03
本文编辑：陈咏竹

(上接第934页；*Continued from Page* 934)

4 Salvatore C, Cerasa A, Castiglioni I, *et al*. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. J Neurosci Methods, 2014, 222: 230-237.
5 Rana B, Juneja A, Saxena M, *et al*. Regions-of-interest based automated diagnosis of Parkinson's disease using T1-weighted MRI. Expert Syst Appl, 2015, 42(9): 4506-4516.
6 Adeli E, Shi Feng, An Le, *et al*. Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. Neuroimage, 2016, 141: 206-219.
7 Peng Bo, Wang Suhong, Zhou Zhiyong, *et al*. A multilevel-ROI-features-based machine learning method for detection of morphometric biomarkers in Parkinson's disease. Neurosci Lett, 2017, 651: 88-94.
8 Shen D, Wu G, Suk H I. Deep learning in medical image analysis. Annu Rev Biomed Eng, 2017, 19: 221.
9 Plis S M, Hjelm D R, Salakhutdinov R, *et al*. Deep learning for neuroimaging: a validation study. Front Neurosci, 2014, 8(8): 229.
10 陈诗慧, 刘维湘, 秦璟, 等. 基于深度学习和医学图像的癌症计算机辅助诊断研究进展. 生物医学工程学杂志, 2017, 34(2): 314-319.
11 Payan A, Montana G. Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. Computer Science, 2015. arXiv: 1502.02506.
12 Suk H I, Lee S W, Shen Dinggang, *et al*. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct, 2015, 220(2): 841-859.
13 Livni R, Shalev-Shwartz S, Shamir O. An algorithm for training polynomial networks. Computer Science, 2013, 26(18): 4748-4750.
14 Shi J, Zheng X, Li Y, *et al*. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE Journal of Biomedical & Health Informatics, 2018, 22(1): 173-183.
15 Shi Jun, Zhou Shichong, Liu Xiao, *et al*. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing, 2016, 194: 87-94.
16 Zhu P, Zhang L, Hu Q, *et al*. Multi-scale patch based collaborative representation for face recognition with margin distribution optimization// European Conference on Computer Vision. 2012, 7572: 822-835.
17 Yang F, Lu H, Yang M H. Robust visual tracking via multiple kernel boosting with affinity constraints. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 242-254.
18 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 2013, 35(8): 1798-1828.

收稿日期：2017-09-14　修回日期：2018-02-28
本文编辑：李蓓兰