第 37 卷 第 3 期
2022 年 6 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 37 No. 3
Jun. 2022

DOI:10.13364/j.issn.1672-6510.20220019

基于拉普拉斯正则化的药物副作用频率预测

王 林，李冰纯，徐显嵛
（天津科技大学人工智能学院，天津 300457）

摘 要：药物风险–效益评价中的一个重要问题是确定药物副作用的频率. 相较于通常的随机对照实验，基于机器学习预测药物副作用频率的方法具有时间短、准确率高的特点，并且可以用来指导对照实验. 现有的计算方法很少考虑“相似的药物具有相似的副作用频率”这一特点，因此预测性能仍有待进一步提高. 本文提出结合拉普拉斯正则化的非负矩阵分解方法，并引入超参数控制未知副作用标签及其预测值的间隔. 计算实验表明，该方法可以有效预测药物的副作用频率，并且还可以预测上市后药物的副作用.

关键词：药物；副作用频率；机器学习；拉普拉斯正则化
中图分类号：TP399 **文献标志码**：A **文章编号**：1672-6510(2022)03-0067-06

Prediction of Frequencies of Drug Side Effects with Laplace Regularization

WANG Lin，LI Bingchun，XU Xianyu
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：An important issue in drug risk-benefit assessment is to determine the frequency of drug side effects. Compared with the usual randomized controlled trials，the method based on machine learning to predict the frequency of drug side effects has the characteristics of short time and high accuracy，and can be used to guide controlled trials. However，existing computational methods rarely take into account the feature that “similar drugs have similar frequency of side effects”，so the prediction performance can be further improved. Therefore a non-negative matrix factorization method combined with Laplace regularization is proposed in this article，and a hyperparameter is also introduced to control the margin between labels and their predicted scores for unknown side effects. Computational experiments show that this method can effectively predict the frequency of drug side effects，and can also predict post-marketing drug side effects.

Key words：drug；side effect frequency；machine learning；Laplace regularization

药物风险–效益评价是对患者用药后得到的治疗效益与风险之间的评价. 在这项评价中，药物副作用频率的估计至关重要[1]. 目前，计算频率的标准方法是随机对照实验，通过对不同分组实施不同的干预措施，得到不同的结果[2]. 这种方法容易受到时间、样本量和熟练度的限制，使药物的一些副作用在临床试验中没有发现，而是在上市多年后被发现[3]. 因此在医疗卫生领域中，药物的副作用仍然是引起其他疾病和死亡的主要原因[4]. 现有的一些预测药物副作用的计算方法[5-7]大多数只能预测副作用存在与否，不能预测副作用的频率，在一定程度上限制了这些方法在药物风险–效益评价中的应用.

Galeano 等[8]提出了利用非负矩阵分解模型（nonnegative matrix factorization，NMF）预测药物的副作用频率，但是该方法对药物副作用关联和频率预测的准确率仍有待提高. 在此基础上，本文提出了一种基于拉普拉斯正则化的药物副作用频率预测模型 DSLR（drug-side effect frequency prediction with

收稿日期：2022–01–26；**修回日期**：2022–03–17
基金项目：天津市教委科研计划项目（2018KJ107）
作者简介：王 林（1982—），男，山西人，副教授，博士，linwang@tust.edu.cn

Laplace regularization)，在非负矩阵分解模型中引入拉普拉斯正则化项，以及控制未知副作用标签及其预测值间隔的超参数. 实验结果和数据分析表明，DSLR 模型不仅能更准确地识别药物的副作用关联，而且能更精确地进行药物副作用频率的预测.

1　数据获取

利用 Galeano 等[8]和 Zhao 等[9]使用的基准数据集验证药物副作用频率预测方法的有效性. 该数据集包括 750 种药物和 994 种副作用，以及来自 SIDER 数据库[10]的 37 071 个已知频率项. 药物副作用依据干预队列，临床试验频率被映射成 5 个频率(f)区间，即 $f<0.01\%$、$0.01\%\leqslant f<0.1\%$、$0.1\%\leqslant f<1\%$、$1\%\leqslant f\leqslant 10\%$ 和 $f>10\%$ 分别定义为罕见、少见、不经常、频繁和非常频繁，并分别用频率值 1、2、3、4、5 表示. 在 37 071 个已知频率项中，罕见、少见、不经常、频繁和非常频繁的占比分别为 3.21%、11.29%、26.92%、47.46% 和 11.12%. 用评级矩阵 $\boldsymbol{M}$ 表示药物和副作用之间的频率，其中矩阵的行和列分别表示药物和副作用，矩阵中的非 0 值表示特定药物–副作用对的已知频率，0 表示未知副作用. 评级矩阵 $\boldsymbol{M}$ 极其稀疏，非零元素仅占 4.97%.

2　计算方法

2.2　药物相似性和副作用相似性的构建

首先，使用开源化学信息 Python 软件包 RDKit，基于拓扑指纹和 Tanimoto 相似度计算任意两种药物之间的化学结构相似度，并表示为矩阵 $\boldsymbol{A}_{\mathrm{w}}^{1}\in\boldsymbol{R}^{n\times n}$，其中 n 为药物的个数. 其次，基于评级矩阵 $\boldsymbol{M}$，计算任意两种药物频率谱($\boldsymbol{M}$ 中的两行)的余弦相似度，并表示为矩阵 $\boldsymbol{A}_{\mathrm{w}}^{2}\in\boldsymbol{R}^{n\times n}$. 最后，取两种相似度的平均值作为药物的相似度，即 $\boldsymbol{A}_{\mathrm{w}}=(\boldsymbol{A}_{\mathrm{w}}^{1}+\boldsymbol{A}_{\mathrm{w}}^{2})/2$.

基于评级矩阵 $\boldsymbol{M}$，计算任意两种副作用频率谱($\boldsymbol{M}$ 中的两列)的余弦相似度，并表示为矩阵 $\boldsymbol{A}_{\mathrm{h}}\in\boldsymbol{R}^{m\times m}$，其中 m 为副作用的个数.

2.2　基于拉普拉斯正则化的优化模型

采用非负矩阵分解模型，将评级矩阵 $\boldsymbol{M}$ 分解为基矩阵 $\boldsymbol{W}\in\boldsymbol{R}^{n\times k}$ 和系数矩阵 $\boldsymbol{H}\in\boldsymbol{R}^{k\times m}$，具体为

$$\min_{\boldsymbol{W},\boldsymbol{H}}\frac{1}{2}\left\|\boldsymbol{M}-\boldsymbol{W}\boldsymbol{H}\right\|_{\mathrm{F}}^{2}\ \mathrm{s.t.}\boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0 \tag{1}$$

其中：$\|\cdot\|_{\mathrm{F}}$ 表示 Frobenius 范数，潜在特征维度 $k=200$. $\boldsymbol{M}$ 中包括非零元素(已知频率)以及零元素(表示未知频率)，并且非零元素相较于零元素，在优化模型中更重要. 将零元素和非零元素的拟合分别赋予不同的权重，即 $\alpha=1$ 和 $\alpha=0.05$，模型变为

$$\min_{\boldsymbol{W},\boldsymbol{H}}\frac{1}{2}\left\|\boldsymbol{I}^{\Omega}\odot(\boldsymbol{M}-\boldsymbol{W}\boldsymbol{H})\right\|_{\mathrm{F}}^{2}+\frac{\alpha}{2}\left\|\boldsymbol{I}^{\mathrm{o}}\odot\boldsymbol{W}\boldsymbol{H}\right\|_{\mathrm{F}}^{2}$$
$$\mathrm{s.t.}\boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0 \tag{2}$$

其中：$\boldsymbol{I}^{\Omega}\in\boldsymbol{R}^{n\times m}$ 是映射矩阵，当 $\boldsymbol{M}$ 中元素为非零时，$\boldsymbol{I}^{\Omega}$ 对应位置为 1，否则 $\boldsymbol{I}^{\Omega}$ 对应位置为 0；$\boldsymbol{I}^{\mathrm{o}}\in\boldsymbol{R}^{n\times m}$ 也是映射矩阵，当 $\boldsymbol{M}$ 中元素为零时，$\boldsymbol{I}^{\mathrm{o}}$ 对应位置为 1，否则 $\boldsymbol{I}^{\mathrm{o}}$ 对应位置为 0；$\odot$ 表示 Hadamard 积. 对于未知的药物–副作用关联($\boldsymbol{M}$ 中的零元素)，引入超参数 ε 控制其预测值和标签 0 的间隔，得到模型为

$$\min_{\boldsymbol{W},\boldsymbol{H}}\frac{1}{2}\left\|\boldsymbol{I}^{\Omega}\odot(\boldsymbol{M}-\boldsymbol{W}\boldsymbol{H})\right\|_{\mathrm{F}}^{2}+\frac{\alpha}{2}\left\|\boldsymbol{I}^{\mathrm{o}}\odot(\varepsilon\boldsymbol{E}-\boldsymbol{W}\boldsymbol{H})\right\|_{\mathrm{F}}^{2}$$
$$\mathrm{s.t.}\boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0 \tag{3}$$

其中：$\boldsymbol{E}$ 表示元素全为 1 的矩阵.

引入拉普拉斯正则化项，即对于基矩阵 $\boldsymbol{W}$，相似的药物对应的行向量也相似；对于系数矩阵 $\boldsymbol{H}$，相似的副作用对应的列向量也相似，从而模型最终变为

$$\begin{aligned}\min_{\boldsymbol{W},\boldsymbol{H}}F(\boldsymbol{W},\boldsymbol{H})=&\frac{1}{2}\left\|\boldsymbol{I}^{\Omega}\odot(\boldsymbol{M}-\boldsymbol{W}\boldsymbol{H})\right\|_{\mathrm{F}}^{2}+\\&\frac{\alpha}{2}\left\|\boldsymbol{I}^{\mathrm{o}}\odot(\varepsilon\boldsymbol{E}-\boldsymbol{W}\boldsymbol{H})\right\|_{\mathrm{F}}^{2}+\\&\frac{\beta}{2}(\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{D}_{\mathrm{w}}-\boldsymbol{A}_{\mathrm{w}})\boldsymbol{W})+\\&\mathrm{tr}(\boldsymbol{H}(\boldsymbol{D}_{\mathrm{h}}-\boldsymbol{A}_{\mathrm{h}})\boldsymbol{H}^{\mathrm{T}}))\\&\mathrm{s.t.}\boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0\end{aligned} \tag{4}$$

其中：β 为超参数，tr() 表示矩阵的迹，$\boldsymbol{D}_{\mathrm{w}}$ 和 $\boldsymbol{D}_{\mathrm{h}}$ 为对角矩阵. $\boldsymbol{D}_{\mathrm{w}}$ 对角线上的第 i 个元素为 $\boldsymbol{D}_{\mathrm{w}}(i)=\sum_{j=1}^{n}\boldsymbol{A}_{\mathrm{w}}(i,j)$，$\boldsymbol{D}_{\mathrm{h}}$ 对角线上的第 i 个元素为 $\boldsymbol{D}_{\mathrm{h}}(i)=\sum_{j=1}^{m}\boldsymbol{A}_{\mathrm{h}}(i,j)$.

2.3　求解算法

采用乘性更新算法求解模型(4). 具体来说，随机初始化 $\boldsymbol{W}$ 和 $\boldsymbol{H}$，并分别用其 Frobenius 范数归一化，进而 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的更新公式为

$$\begin{cases}\boldsymbol{W}=\boldsymbol{W}_{0}\odot\sqrt{\dfrac{(\boldsymbol{M}+\alpha\varepsilon\boldsymbol{I}^{\mathrm{o}})\boldsymbol{H}_{0}^{\mathrm{T}}+\beta\boldsymbol{A}_{\mathrm{w}}\boldsymbol{W}_{0}}{(\boldsymbol{I}^{\Omega}\odot(\boldsymbol{W}_{0}\boldsymbol{H}_{0})+\alpha\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{W}_{0}\boldsymbol{H}_{0}))\boldsymbol{H}_{0}^{\mathrm{T}}+\beta\boldsymbol{D}_{\mathrm{w}}\boldsymbol{W}_{0}}}\\\boldsymbol{H}=\boldsymbol{H}_{0}\odot\sqrt{\dfrac{\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{M}+\alpha\varepsilon\boldsymbol{I}^{\mathrm{o}})+\beta\boldsymbol{H}_{0}\boldsymbol{A}_{\mathrm{h}}}{\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{I}^{\Omega}\odot(\boldsymbol{W}\boldsymbol{H}_{0})+\alpha\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{W}\boldsymbol{H}_{0}))+\beta\boldsymbol{H}_{0}\boldsymbol{D}_{\mathrm{h}}}}\end{cases} \tag{5}$$

其中：$\boldsymbol{W}_{0}$ 和 $\boldsymbol{H}_{0}$ 为更新前的矩阵，$\boldsymbol{W}$ 和 $\boldsymbol{H}$ 为更新后的矩阵. 基于更新公式(5)，模型(4)的目标函数是单调

下降的，从而可以保证算法的收敛性. 设置最大迭代次数为 1 000，并且当前后两次迭代目标函数的下降值小于设定阈值时，停止迭代.

算法执行前，首先运用 $\boldsymbol{M}/5$ 将评级矩阵 $\boldsymbol{M}$ 归一化，然后采用上述乘性更新算法得到 $\boldsymbol{W}$ 和 $\boldsymbol{H}$，进而令 $\boldsymbol{P}=\boldsymbol{WH}$，最后通过 $\boldsymbol{P}\times 5$ 得到最终的预测矩阵.

2.4　收敛性分析

根据约束最优化理论[11]，当目标函数收敛时，最优解满足的 Karush-Kuhn-Tucker（KKT）互补条件为

$$\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{W}}\right)_{ip}\boldsymbol{W}_{ip}=0,\left(\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{H}}\right)_{pj}\boldsymbol{H}_{pj}=0 \tag{6}$$

其中：$i\in\{1,\cdots,n\}$，表示 n 个药物的索引；$j\in\{1,\cdots,m\}$，表示 m 个副作用的索引；$p\in\{1,\cdots,k\}$，表示 k 个潜在特征的索引.

式（4）的梯度可以写成矩阵形式，为

$$\begin{cases}\dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{W}}=-\left(\boldsymbol{M}-\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})\right)\boldsymbol{H}^{\mathrm{T}}-\\ \qquad \alpha\left(\varepsilon\boldsymbol{I}^{\mathrm{o}}-\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)\boldsymbol{H}^{\mathrm{T}}+\beta\boldsymbol{D}_{\mathrm{w}}\boldsymbol{W}-\beta\boldsymbol{A}_{\mathrm{w}}\boldsymbol{W}\\ \dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{H}}=-\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{M}-\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})\right)-\boldsymbol{W}^{\mathrm{T}}\alpha\left(\varepsilon\boldsymbol{I}^{\mathrm{o}}-\\ \qquad \boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)+\beta\boldsymbol{HD}_{\mathrm{h}}-\beta\boldsymbol{HA}_{\mathrm{h}}\end{cases} \tag{7}$$

当 $\boldsymbol{W}=\boldsymbol{W}^{*}$ 且 $\boldsymbol{H}=\boldsymbol{H}^{*}$ 使模型（4）取得局部极小值时，必须满足式（6）中的 KKT 互补条件，其中 $\boldsymbol{W}^{*}$ 和 $\boldsymbol{H}^{*}$ 表示局部最优解. 将式（7）代入式（6），得

$$\begin{cases}\left(\begin{array}{l}-\left(\boldsymbol{M}-\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})\right)\boldsymbol{H}^{\mathrm{T}}-\\ \alpha\left(\varepsilon\boldsymbol{I}^{\mathrm{o}}-\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)\boldsymbol{H}^{\mathrm{T}}+\\ \beta\boldsymbol{D}_{\mathrm{w}}\boldsymbol{W}-\beta\boldsymbol{A}_{\mathrm{w}}\boldsymbol{W}\end{array}\right)_{ip}\boldsymbol{W}_{ip}^{2}=0\\ \left(\begin{array}{l}-\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{M}-\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})\right)-\\ \boldsymbol{W}^{\mathrm{T}}\alpha\left(\varepsilon\boldsymbol{I}^{\mathrm{o}}-\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)+\\ \beta\boldsymbol{HD}_{\mathrm{h}}-\beta\boldsymbol{HA}_{\mathrm{h}}\end{array}\right)_{pj}\boldsymbol{H}_{pj}^{2}=0\end{cases} \tag{8}$$

将式（8）重新整理，得

$$\begin{cases}-\left(\left(\boldsymbol{M}+\alpha\varepsilon\boldsymbol{I}^{\mathrm{o}}\right)\boldsymbol{H}^{\mathrm{T}}+\beta\boldsymbol{A}_{\mathrm{w}}\boldsymbol{W}\right)_{ip}\boldsymbol{W}_{ip}^{2}+\\ \left(\left(\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})+\alpha\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)\boldsymbol{H}^{\mathrm{T}}+\beta\boldsymbol{D}_{\mathrm{w}}\boldsymbol{W}\right)_{ip}\boldsymbol{W}_{ip}^{2}=0\\ -\left(\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{M}+\alpha\varepsilon\boldsymbol{I}^{\mathrm{o}}\right)+\beta\boldsymbol{HA}_{\mathrm{h}}\right)_{pj}\boldsymbol{H}_{pj}^{2}+\\ \left(\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{I}^{\Omega}\odot(\boldsymbol{WH})+\alpha\boldsymbol{I}^{\mathrm{o}}\odot(\boldsymbol{WH})\right)+\beta\boldsymbol{HD}_{\mathrm{h}}\right)_{pj}\boldsymbol{H}_{pj}^{2}=0\end{cases} \tag{9}$$

结合式（9）不难看出，$\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的更新公式（5）满足 KKT 互补条件，从而基于式（5）则模型（4）收敛到局部最小值.

2.5　预测性能的度量

预测模型的准确性从两个方面衡量，即识别药物副作用关联的性能和频率预测的性能. 对于二分类问题，可以将实例（药物–副作用对）分为正例（有关联）或负例（未知关联）. 进行预测时，会出现以下 4 种情况：True Positive（TP），实例是正例并被预测为正例；False Positive（FP），实例是负例并被预测为正例；False Negative（FN），实例是正例并被预测为负例；True Negative（TN），实例是负例并被预测为负例.

准确率（Accuracy）、精确率（Precision）、召回率（Recall）的计算式为

$$\text{Accuracy}=(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \tag{10}$$

$$\text{Precision}=\text{TP}/(\text{TP}+\text{FP}) \tag{11}$$

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

此外两个常用的指标，即 Precision-Recall（PR）曲线下面积（area under the precision-recall curve，AUPR）以及接受者操作特征曲线（receiver operating characteristic curve，ROC）下面积（area under curve，AUC）也用来评价关联性能.

对于每个指标，首先计算测试集上每种药物的指标值. 对于每种给定的药物，其在测试集中具有已知频率的副作用和其在评级矩阵 $\boldsymbol{M}$ 中的未知副作用分别被视为正例和负例，然后将所有药物的平均指标值作为结果.

关于频率预测，使用 Spearman 相关系数（Spearman's correlation coefficient，SCC）和均方根误差（root mean square error，RMSE）作为评价指标，SCC 和 RMSE 的计算式为

$$\text{SCC}=\frac{\sum_{d,e}(r(P_{d,e})-\overline{r(P_{d,e})})(r(M_{d,e})-\overline{r(M_{d,e})})}{\sqrt{\sum_{d,e}(r(P_{d,e})-\overline{r(P_{d,e})})^{2}\sum_{d,e}(r(M_{d,e})-\overline{r(M_{d,e})})^{2}}} \tag{13}$$

$$\text{RMSE}=\sqrt{\frac{\sum_{d,e}(P_{d,e}-M_{d,e})^{2}}{t}} \tag{14}$$

其中：d 和 e 分别表示药物和副作用的遍历，$P_{d,e}$ 和 $M_{d,e}$ 分别表示药物–副作用对的预测频率和已知频率，$r()$ 表示等级转换，t 表示已知频率的药物–副作用对的总数.

3　计算结果与讨论

3.1　化学结构相似的药物有相似的副作用频率

使用开源化学信息 Python 软件包 RDKit，基于

拓扑指纹和 Tanimoto 相似度计算任意两个药物之间的化学结构相似性. 对于 280 875 个药物对，其化学结构相似性的中位数为 0.24，将相似性≤0.24 的药物对定义为化学结构低相似度对，将相似性＞0.24 的药物对定义为化学结构高相似度对.

对于 280 875 个药物对，计算其副作用频率相似度，即对于任意两个药物，基于其副作用频率谱（评级矩阵 $\boldsymbol{M}$ 中的两行），利用余弦相似度进行计算. 药物对关于副作用频率余弦相似度的箱线图如图 1 所示. 图 1 给出了化学结构低相似度对和高相似度对的副作用频率相似度分布的箱线图，相对于化学结构低相似度对，化学结构高相似度对具有更大的副作用频率相似度（单边 Wilcoxon 秩和检验 $P = 5.85 \times 10^{-59}$）.

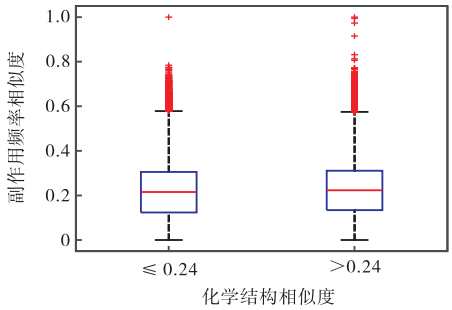


图 1　药物对关于副作用频率余弦相似度的箱线图

Fig. 1　Box plots of drug pairs with respect to the cosine similarity between their side effect frequency profiles

3.2　10 折交叉验证

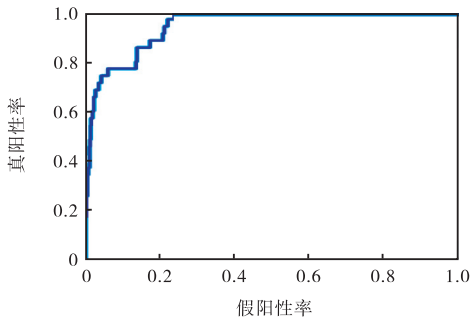
在数据集中，所有已知药物–副作用对的频率（共计 37 071 个）被随机均匀地分成 10 折. 数据集的其中一折设置为测试集，其余 9 折则作为训练集，并将每一折测试集的平均指标值作为最终结果. 选择现有的副作用频率预测模型 NMF[8]和 MGPred（prediction using a graph attention network to integrate multi-view data）[9]作为对比，验证本文模型 DSLR 的有效性. 同时，考虑建模副作用频率预测问题为推荐系统，采用基于图神经网络的模型（inductive graph-based matrix completion，IGMC）[12]求解. 基于 10 折交叉验证的比较结果见表 1. 由表 1 可知：DSLR 模型的 AUC、AUPR 明显优于其他 3 个模型，这表明 DSLR 模型可以对药物副作用关联进行更好地预测；对于评价频率预测性能的指标，DSLR 模型的 SCC 和 RMSE 明显优于 NMF 模型，但逊于 MGPred 模型和 IGMC 模型. MGPred 和 IGMC 这两个模型的 AUC 较低，表明其不能准确地预测药物–副作用关联，因此虽然其 SCC 和 RMSE 更优，但在实际使用中容易引入假阳性，即未知副作用大多数被预测为有药物–副作用关联.

使用 DSLR 模型对单个药物氟伏沙明进行研究. 对于 10 折交叉验证中的 1 折，测试集中氟伏沙明的已知副作用共 35 个（正例），未知副作用共 694 个（负例）. 选取与正例等量的负例，计算得出 Accuracy = 0.614，Precision = 0.565，Recall = 1.0. 该药物对于 729 个副作用预测结果的 ROC 曲线（AUC = 0.948）和 PR 曲线（AUPR = 0.559）见图 2.

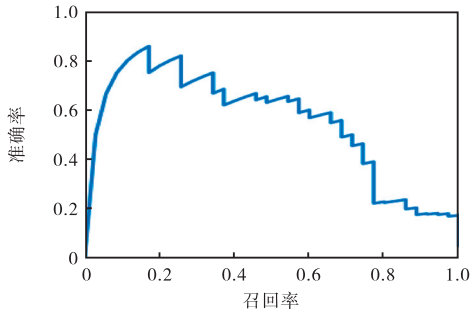
表 1　基于 10 折交叉验证的比较结果

Tab. 1　Comparison results based on 10-fold cross validation

模型	AUPR	AUC	SCC	RMSE
NMF	0.220	0.907	0.494	1.285
MGPred	0.135	0.771	0.723	0.663
IGMC	0.119	0.745	0.750	0.618
DSLR	0.289	0.922	0.539	1.114



（a）ROC 曲线



（b）PR 曲线

图 2　药物氟伏沙明副作用预测的 ROC 曲线和 PR 曲线

Fig. 2　ROC curve and PR curve for the prediction of the side effects of the drug fluvoxamine

为了进行频率类别预测，使用 10 折交叉验证期间从测试集得到的预测值，收集了所有已知副作用的频率类别及其对应的预测值. 对于未知副作用，基于 10 折交叉验证中的 1 折，得到未知副作用的预测值. 对于未知副作用及已知副作用的每个频率类别，采用核密度估计方法得到其预测值的概率密度函数

(probability density function，PDF)．每一频率类别预测值的概率密度函数如图 3 所示，其中频率 0～5 分别对应副作用频率类别为未知副作用、罕见、少见、不经常、频繁和非常频繁．

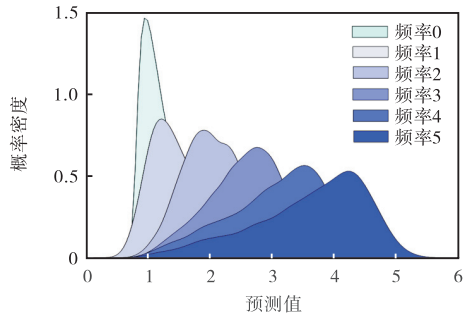


图 3　每一频率类别预测值的概率密度函数

Fig. 3　PDF of predicted values for each frequency category

根据概率密度函数和最大似然法确定分类决策的边界，得到相邻频率的边界阈值分别为 1.15、1.65、2.35、3.05 和 3.85(图 3)．对于每一个真实频率类别中的所有副作用，可以得到其预测频率类别．每一频率类别的准确率见表 2．表 2 给出了预测为各个类别的副作用所占的百分比，对于频繁(频率 = 4)副作用(占总数的 47.46%)中的 41.89% 被正确预测，79.67% 被预测为不经常(频率 = 3)、频繁或非常频繁(频率 = 5)．

表 2　每一频率类别的准确率

Tab. 2　Accuracy for each frequency category

		预测频率					
		0	1	2	3	4	5
真实频率	0	48.5	32.6	13.87	3.85	1.06	0.11
	1	24.87	37.98	25.8	8.91	2.35	0.08
	2	2.7	20.19	51.62	23.36	2.01	0.12
	3	1.12	7.37	26.45	44.63	19.37	1.06
	4	1.01	4.89	14.43	24.78	41.89	13.0
	5	0.61	2.57	8.39	13.66	28.47	46.3

进一步定义精确类和邻居类两个概念．精确类是被预测为自身真实频率的类别，如真实频率为 1 的副作用被预测为频率类别 1．邻居类是指被预测为自身和其邻居真实频率的类别，如真实频率为 1 的副作用被预测为频率类别 1 和 2，真实频率为 2 的副作用被预测为频率类别 1、2 和 3．

本研究对单个药物盐酸罗匹尼罗进行了分析，该药物共有 396 个副作用，频率为 1、2、3、4 和 5 的副作用个数分别为 0、17、167、209 和 3．该药物频率为 2、3、4 和 5 的精确类准确率分别为 11.76%、28.74%、30.62% 和 33.33%，邻居类准确率分别为 41.17%、68.26%、84.21% 和 66.66%．

3.3　消融实验

在引入拉普拉斯正则化项以及控制未知副作用标签和其预测值间隔的超参数ε后，验证 DSLR 模型在预测药物副作用频率方面的优越性(表 3)．对于给定的基准数据集，引入拉普拉斯正则化项对模型预测药物–副作用关联的性能有明显提升；引入超参数ε，在 AUC 相对稳健的情况下，RMSE 显著降低，表明其能更精确地进行频率预测．因此，当拉普拉斯正则化项的权重参数 $\beta = 0.01$、间隔$\varepsilon = 0.195$ 时，AUC = 0.922，RMSE = 1.114，DSLR 模型的预测性能最好．

表 3　消融实验的比较结果

Tab. 3　Comparison results of ablation experiments

模型	AUPR	AUC	SCC	RMSE
$\beta = 0, \varepsilon = 0$	0.198	0.885	0.278	2.095
$\beta = 0.01, \varepsilon = 0$	0.301	0.928	0.465	1.454
$\beta = 0, \varepsilon = 0.195$	0.243	0.861	0.334	1.695
DSLR	0.289	0.922	0.539	1.114

3.4　上市后副作用预测

对于基准数据集的 750 种药物和 994 种副作用，本研究发现评级矩阵 $\boldsymbol{M}$ 的未知副作用中，有 9 288 种药物–副作用关联在 SIDER 数据库中被标记为“上市后”(以下简称上市后副作用)．这些上市后副作用由于在临床试验中并没有发现，被认为频率为 1，即罕见的副作用[13]．使用 $\boldsymbol{M}$ 中所有已知频率类别(频率＞0)作为训练集训练模型，然后对上市后副作用进行预测．图 4 给出了未知副作用(频率 = 0)和上市后副作用预测值的 PDF，以及基于 10 折交叉验证 $\boldsymbol{M}$ 中罕见(频率 = 1)副作用预测值的 PDF．结果表明：对于 9 288 种上市后副作用，有 31.52% 被正确地预测为罕见，62.34% 被预测为罕见或少见(频率 = 2)，82.82% 被识别为有药物–副作用关联，说明 DSLR 模型对上市后副作用有较好的预测能力．

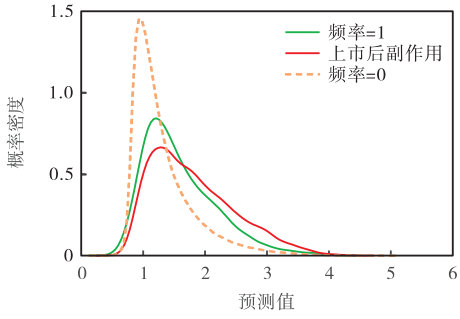


图 4　频率 = 1、频率 = 0 及上市后副作用的概率密度函数

Fig. 4　PDF of predicted values for frequency = 1，frequency = 0 and post-marketing side effects

本研究随机选取了药物舒尼替尼，在 SIDER 数据库中该药物有 51 个副作用在上市后被发现. 对于这些副作用，预测结果表明 86.27% 被识别为有药物–副作用关联，其中 54.9% 被识别为罕见（频率 = 1），21.57% 被识别为少见（频率 = 2），9.8% 被识别为不经常（频率 = 3）.

4 结　语

本文提出了一种预测药物副作用频率的机器学习模型 DSLR. 基于基准数据集，DSLR 模型将药物之间的化学结构相似度和药物频率谱的余弦相似度的平均值作为药物的相似度，副作用频率谱的余弦相似度作为副作用的相似度，采用基于拉普拉斯正则化的非负矩阵分解模型，并引入超参数控制未知副作用标签及其预测值的间隔. 结果表明，DSLR 模型不仅能准确预测药物副作用发生的频率，并且能够对上市后药物副作用进行预测，这有助于指导药物风险–效益评价.

参考文献：

[1] GODAT S，FOURNIER N，SAFRONEEVA E，et al. Frequency and type of drug-related side effects necessitating treatment discontinuation in the Swiss Inflammatory Bowel Disease Cohort[J]. European journal of gastroenterology & hepatology，2018，30(6)：612–620.

[2] CONCATO J，SHAH N，HORWITZ R I. Randomized，controlled trials，observational studies，and the hierarchy of research designs[J]. The New England journal of medicine，2000，342(25)：1887–1892.

[3] BANDA J M，EVANS L，VANGURI R S，et al. A curated and standardized adverse drug event resource to accelerate drug safety research[J]. Scientific data，2016，3(1)：160026.

[4] PIRMOHAMED M，JAMES S，MEAKIN S，et al. Adverse drug reactions as cause of admission to hospital：prospective analysis of 18 820 patients[J]. British medical journal，2004，329(7456)：15–19.

[5] CAMI A，ARNOLD A，MANZI S，et al. Predicting adverse drug events using pharmacological network models[J]. Science translational medicine，2011，3(114)：114–127.

[6] WANG Z，CLARK N R，MA'AYAN A. Drug-induced adverse events prediction with the LINCS L1000 data[J]. Bioinformatics，2016，32(15)：2338–2345.

[7] CAKIR A，TUNCER M，TAYMAZ-NIKEREL H，et al. Side effect prediction based on drug-induced gene expression profiles and random forest with iterative feature selection[J]. The pharmacogenomics journal，2021，21：673–681.

[8] GALEANO D，LI S，GERSTEIN M，et al. Predicting the frequencies of drug side effects[J]. Nature communications，2020，11(1)：4575.

[9] ZHAO H，ZHANG K，LI Y，et al. A novel graph attention model for predicting frequencies of drug-side effects from multi-view data[J]. Briefings in bioinformatics，2021，22(6)：239.

[10] KUHN M，LETUNIC I，JENSEN L J，et al. The SIDER database of drugs and side effects[J]. Nucleic acids research，2016，44(1)：1075–1079.

[11] LI T，DING C. The relationships among various nonnegative matrix factorization methods for clustering[C]//IEEE. Sixth International Conference on Data Mining(ICDM'06). New York：IEEE，2006：4053063.

[12] ZHANG M，CHEN Y. Inductive matrix completion based on graph neural networks[EB/OL]. [2022–01–25]. https://arxiv.org/abs/1904.12058.

[13] TATONETTI N P，YE P P，DANESHJOU R，et al. Data-driven prediction of drug effects and interactions[J]. Science translational medicine，2012，4(125)：125–131.

责任编辑：郎婧