融合知识图谱与深度学习的药物发现方法

桑盛田[1] 杨志豪[1] 刘晓霞[1] 王 磊[2] 赵 迪[1] 林鸿飞[1] 王 健[1]

摘 要 海量增长的生物医学文献给文献挖掘技术带来巨大挑战. 文中提出融合知识图谱与深度学习的药物发现方法,从已发表的文献中挖掘疾病的潜在治疗药物. 首先抽取生物医学文献中实体间的关系,构造生物医学知识图谱,再通过知识图谱嵌入方法将知识图谱中的实体和关系转化为低维连续的向量,最后使用已知的药物疾病关系数据训练基于循环神经网络的药物发现模型. 实验表明,文中方法不仅可以有效找到疾病的候选药物,还能提供相应的药物作用机制.

关键词 数据挖掘, 生物医学知识图谱, 深度学习, 循环神经网络

引用格式 桑盛田,杨志豪,刘晓霞,王 磊,赵 迪,林鸿飞,王 健. 融合知识图谱与深度学习的药物发现方法. 模式识别与人工智能, 2018, 31(12): 1103-1110.

DOI 10.16451/j.cnki.issn1003-6059.201812005 **中图法分类号** TP 391

A Method Combining Knowledge Graph and Deep Learning for Drug Discovery

SANG Shengtian[1], YANG Zhihao[1], LIU Xiaoxia[1], WANG Lei[2], ZHAO Di[1], LIN Hongfei[1], WANG Jian[1]

ABSTRACT The massive growing amount of biomedical literature brings huge challenges for data mining. In this paper, a method combining knowledge graph and deep learning is proposed to discover potential therapeutic drugs for disease of interest. Firstly, a biomedical knowledge graph is constructed with the relations extracted from biomedical literature. Then, the entities and relations of the knowledge graph are converted into low dimension continuous embeddings by knowledge graph embedding method. Finally, a recurrent neural network based drug discovery model is trained by using the known drug-disease related associations. The experimental results show that the proposed method can discover drugs for diseases and provide the drug mechanism of action.

Key Words Data Mining, Biomedical Knowledge Graph, Deep Learning, Recurrent Neural Network

Citation SANG S T, YANG Z H, LIU X X, WANG L, ZHAO D, LIN H F, WANG J. A Method Combining Knowledge Graph and Deep Learning for Drug Discovery. Pattern Recognition and Artificial Intelligence, 2018, 31(12): 1103-1110.

药物发现(Drug Discovery)是一个周期漫长且代价昂贵的过程. 开发一款新药平均需要 14 年和 18 亿美金[1]. 相反地,从文献中挖掘新的药物是一个周期相对较短且经济的方法. 目前,PubMed 数据库收录超过 2 400 万篇生物医学摘要,通过挖掘这些海量的医学文献可以找到某些疾病潜在的治疗方法[2]. 例如在 1986 年以前雷诺士病(Raynaud Disea-

收稿日期:2018-10-25;录用日期:2018-12-12
Manuscript received October 25, 2018;
accepted December 12, 2018
国家十三五重点研发计划项目(No. 2016YFC0901902)、国家自然科学基金项目(No. 61272373)资助
Supported by National Key Research and Development Project of China(No. 2016YFC0901902), National Natural Science Foundation of China(No. 61272373)
本文责任编委 马少平
Recommended by Associate Editor MA Shaoping

1. 大连理工大学 计算机科学与技术学院 大连 116024
2. 中国人民解放军军事医学科学院 卫生勤务与血液研究所 北京 100850
1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024
2. Institute of Health Service and Blood Research, Academy of Military Medical Sciences, Beijing 100850

se)是一种无法治愈的末梢动脉痉挛性疾病. Swanson 通过阅读一部分医学文献发现“雷诺士病患者存在血液和血管相关的生理改变,如血黏度和血小板凝集度升高,血管收缩等现象”. 同时他通过阅读另一部分文献又发现“鱼油及其活性成分可降低血黏度和血小板凝集度,并引起血管舒张”. 通过人工阅读文献进而推理潜在药物的方法虽然可行,但泛化能力较弱,无法处理海量的生物医学文献. 因此,研究人员提出一系列自动的医学文献挖掘方法.

基于共现的方法(Co-occurrence Based Method)通过发现和药物疾病都共现的实体以推测药物和疾病的关系. 在此基础上,基于关系的方法(Relation-Based Techniques)使用实体间明确的关系推理治疗疾病的候选药物. Cohen 等[3]使用自然语言处理技术(Natural Language Processing)从生物医学文献中抽取实体间的关系,再使用这些关系推理疾病的潜在治疗方法. 更进一步,Ahlers 等[4]通过定义语义发现模板(Discovery Pattern)的方法,更准确挖掘药物疾病间的关系. 最近,研究人员提出一系列基于语义图的文献挖掘方法. Hristowski 等[5]提出基于图的语义路径聚类算法,通过语义路径的聚类发现并解释药物和疾病间的潜在联系.

上述挖掘方法在生物医学领域已取得一定成就,但其仍存在缺陷. 基于共现的方法准确率较低,无法解释实体间的联系;基于关系的方法主要解决短路径推理问题,无法推理实体间较长的关系;基于语义发现模板的方法具有较高的准确率,但语义模板的定义需要大量的人工参与;基于图的方法主要用于解释药物疾病间的关系,无法发现疾病的潜在治疗药物.

除上述方法外,学者们将一系列基于机器学习和深度学习的方法用于药物发现的领域中[6]. 这些方法主要使用现有的数据库作为数据源,并未考虑生物医学文献中的数据. 本文提出融合知识图谱与深度学习的基于文献挖掘的药物发现方法(Biomedical Knowledge Graph Embedding Based Deep Learning Method, GrEDeL). 首先使用生物医学文摘中的实体关系构造生物医学知识图谱(SemRep Based Knowledge Graph, SemKG). 然后通过知识图谱嵌入(Knowledge Graph Embedding)将 SemKG 中的实体和关系转化为低维连续的向量. 再使用“药物-靶标-疾病”数据训练双向长短记忆网络(Bidirectional Long Short-Term Memory Networks, BLSTM). 最后使用已训练好的深度学习模型结合知识图谱对新的疾病进行药物挖掘. 实验表明本文方法可以有效发现疾病的潜在治疗药物,并且提供相应的药物作用机制(Mechanism of Action).

1 融合知识图谱与深度学习的药物发现方法

本文方法分为三部分:知识图谱构建,知识图谱嵌入和神经网络模型训练. 原理框图如图 1 所示.

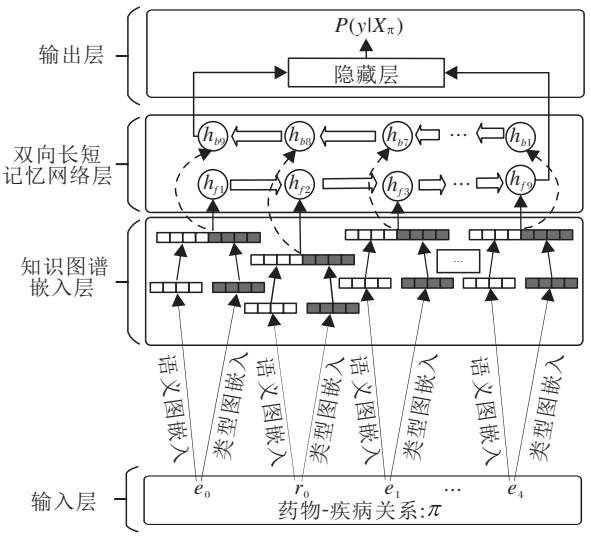


图 1 本文方法原理框图

Fig. 1 Framework of the proposed method

1.1 生物医学知识图谱构建

首先使用关系抽取工具 SemRep[7]从生物医学文摘中抽取实体关系. 如 SemRep 从“We used hemofil-tration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia”中可以抽取出四个实体关系:

1. Hemofiltration | topp TREATS Patients | podg
2. Digoxin overdose | inpo
   PROCESS_OF Patients | podg
3. Hyperkalemia | patf
   COMPLICATES Digoxin overdose | inpo
4. Hemofiltration | topp
   TREATS(INFER) Digoxin overdose | inpo

其中‘|’右边表示该实体的 UMLS 语义类型,如 topp 表示 Therapeutic or Preventive Procedure.

本文利用 SemRep 抽取实体关系构造知识图谱 SemKG. 知识图谱表示为

$$G_{KG}=(E\cup\phi(E),R\cup\phi(R)).$$

其中 $E=\{e_1,e_2,\cdots,e_N\}$ 表示知识图谱中的 N 个实

体,$R=\{r_1,r_2,\cdots,r_M\}$ 表示实体间的关系. $T=\{t_1,t_2,\cdots,t_K\}$ 表示语义类型(UMLS 语义类型). 每个实体 e 或关系 r 可以通过关系映射函数 ϕ 对应语义类型,如 $\phi(e)\rightarrow T_e$ 得到实体类型,$\phi(r)\rightarrow T_r$ 对应关系类型. 图 2 为 SemKG 的示意图.

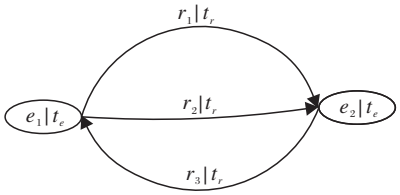


图 2　SemKG 示意图

Fig. 2　Sketch map of SemKG

1.2　训练数据集构造

知识图谱中的路径 π 定义为一个连续实体和关系序列 $e_0r_0e_1r_1e_2r_2\cdots$ 对于一例标准数据 $drug_i$-$target_i$-$disease_i$(已知该药物 $drug_i$ 通过靶标 $target_i$ 作用于 $disease_i$),通过路径搜索算法得到数据集

$$\pi^l=\rho(drug_i\rightarrow disease_i;target_i,l),$$

π^l 表示在知识图谱中以 $drug_i$ 为起点,$disease_i$ 为终点,穿过节点 $target_i$ 的长度为 l 的所有路径. 进一步构造长度为 2 到 l 的路径集合 $l=\{\pi^2,\pi^3,\cdots,\pi^l\}$ 作为本实验的正例集合. 类似地,实验中的负例集合通过三元组 $drug_j'$-$target_j'$-$disease_j'$获得. 该三元组表示药物 $drug_j'$对疾病 $disease_j'$没有治疗作用,或 $target_j'$不是药物 $drug_j'$的靶标.

1.3　基于知识图谱嵌入的双向长短记忆网络药物发现模型

给定一条路径 $\pi_i^l=e_0r_0e_1r_1\cdots r_{l-1}e_l$,其中,$e_0$ 表示某种药物,e_l 表示某种疾病. 药物发现模型的目标是预测该药物治疗该疾病的概率:

$$p(y\mid\pi_i^l)=D(g(\pi_i^l),\theta),$$

其中,$D(\cdot)$ 表示任意具有参数 θ 的判别模型,$g(\cdot)$ 表示特征抽取函数.

如 1.2 节所示,本文药物发现模型的输入层为任意路径 $\pi_i^l=e_0r_0e_1r_1\cdots r_{l-1}e_l$,其中,$e$ 为实体,r 为两个实体间的关系.

在知识图谱嵌入层,π_i^l 中的每个元素 x_i(x_i 为 e 或 r)被转化为向量表示. 如图3所示,首先知识图谱

$$G_{KG}=(E\cup\phi(E),R\cup\phi(R))$$

同时转化成语义图(Semantic Graph)和类型图(Type Graph). 语义图 $G_{SG}=(E,R)$ 只包含实体和实体间的关系. 类型图 $G_{TG}=(\phi(E),\phi(R))$ 只包含实体和关系对应的语义类型. 利用翻译嵌入(Translating Embedding, TransE)[8] 对 G_{SG} 和 G_{TG} 分别进行图嵌入,因此 π_i^l 中的每个元素 x_i 被转化为向量表示:

$$x_i=g(x_i)_{SG}\triangleright\triangleleft g(x_i)_{TG},$$

其中,$g(\cdot)$ 表示知识图谱嵌入方法,符号 $\triangleright\triangleleft$ 表示向量的拼接操作. 通过最小化损失函数得到知识图谱的嵌入表示(以语义图为例):

$$L=\sum_{(e_1,r,e_2)\in S}\sum_{(e_{1'},r,e_{2'})\in S'}[\gamma+d(\boldsymbol{e}_1+\boldsymbol{r},\boldsymbol{e}_2)-d(\boldsymbol{e}_{1'}+\boldsymbol{r},\boldsymbol{e}_{2'})]_+.$$

其中:黑体表示对应元素的向量表示,例如 $\boldsymbol{e}_1$ 为实体 e_1 的向量;$[x]_+$ 表示 x 大于零的部分;d 为 L_1 范式,$\gamma>0$. 正例训练集 $S_{(e_1,r,e_2)}$ 包含所有语义图中的三元组 (e_1,r,e_2),负例集合 S' 由替换正例集合 $S_{(e_1,r,e_2)}$ 三元组的实体 e_1 或 e_2 得到(该步骤需要确保 S' 中不包括 $S_{(e_1,r,e_2)}$ 中的正例数据). 构造学习图嵌入的数据如下:

$$S'_{(e_1,r,e_2)}=\{(e_{1'},r,e_2)\mid e_{1'}\in E\}\cup\{(e_1+r,e_{2'})\mid e_{2'}\in E\}.$$

本节采用随机梯度下降法(Stochastic Gradient Descent, SGD)得到最终的图嵌入表示. 类型图的图嵌入学习过程和语义图相同,本节使用的图嵌入方法的复杂度为 $O(n_ek+n_rk)$,其中,n_e、n_r 分别为知识图谱中实体数量、关系数量,k 为图嵌入向量的维度.

最终,如图 3 所示,知识图谱嵌入层将 π_i^l 中的每个元素被转化为 $L_{SG}+L_{TG}$ 长度的向量,最后 π_i^l 转化为一个 $(L_{SG}+L_{TG})\times l$ 的矩阵 $\boldsymbol{X}$:

$$X_{\pi_i^l}=\bigcup_{x_i\in\pi_i^l}x_i.$$

本文提出使用基于循环神经网络模型(Recurrent Neural Network, RNN)的双向长短记忆网络(BLSTM)预测药物-靶标-疾病关系. 包含输入层 x_t,隐藏层(h_t)和输出层(y_t)的 LSTM 构造如下:

$$\begin{aligned}i_t&=\sigma(W_{xi}x_t+W_{hi}h_{t-1}+b_i),\\f_t&=\sigma(W_{xf}x_t+W_{hf}h_{t-1}+b_f),\\o_t&=\sigma(W_{xo}x_t+W_{ho}h_{t-1}+b_o),\\g_t&=\tanh(W_{xg}x_t+W_{hg}h_{t-1}+b_g),\\c_t&=f\odot c_{t-1}+i_t\odot g_t,\\h_t&=o\odot\tanh(c_t).\end{aligned}$$

其中:σ 为逻辑回归函数(Logistic Sigmoid Function),$\boldsymbol{i},\boldsymbol{f},\boldsymbol{o}$ 和 $\boldsymbol{c}$ 分别表示和隐层向量 h 维数相同的输入门(Input Gate)向量、忘记门(Forget Gate)向量、输出门(Output Gate)向量和细胞(Cell)激活向量;W_* 和 b_* 为可训练的参数,$\odot$ 为位乘操作,c_0 为输入 π_i^l 中的 e_0 的向量表示(e_0 表示某种药物). 传统长短记忆网络(Long Short-Term Memory, LSTM)的一个缺

点是只使用前文信息,而 π_i^l 中实体的顺序影响药物和疾病关系. 因此本文采用 BLSTM 设计药物发现模型. 如图 1 所示,BLSTM 利用两个隐层在不同方向处理数据:

$$h_{ft} = H(W_{xh_f} x_t + W_{h_f h_f} h_{f_{t-1}} + b_{h_f}),$$

$$h_{bt} = H(W_{xh_b} x_t + W_{h_b h_b} h_{b_{t-1}} + b_{h_b}),$$

其中 h_f 和 h_b 分别为前向层和后向层的隐层. 最后,输入模型中的药物-疾病关系概率为

$$p(y \mid X) = \sigma(W_{h_f z} h_t + W_{h_b z} h_t + b_z),$$

其中 $W_{h_f z}$、$W_{h_b z}$、b_z 为待训练的参数. 为了防止训练模型产生过拟合,在 BLSTM 中的非循环部分加入 dropout. 最终,通过反向传播算法(Back Propagation Through Time, BPTT)优化交叉熵损失函数(Cross Entropy Loss Function)$L(\theta)$:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y\ln(p(y \mid X_i)) + (1-y)\ln(1-p(y \mid X_i))).$$

GrEDeL 的目标是寻找给定疾病 $disease_i$ 的潜在治疗药物. 过程如下:对每个候选药物 $drug_{\text{potential}}$ 构造路径集合

$$\Pi_{\text{potential}} = \{\pi^2, \pi^3, \cdots, \pi^l\},$$

其中

$$\pi^l = \rho(drug_{\text{potential}} \rightarrow disease_i; target_{\text{potential}}, l)$$

表示在知识图谱中起点为候选药物 $drug_{\text{potential}}$、终点为 $disease_i$、穿过靶标 $target_{\text{potential}}$ 的所有长度为 l 的路径. GrEDeL 对数据集 $\Pi_{\text{potential}}$ 中的每个数据打分,最终该候选药物 $drug_{\text{potential}}$ 的分数为

$$score(drug_{\text{potential}}) = \max_{\pi_i \in \Pi_{\text{potential}}} D(g(\pi_i), \theta).$$

需要注意的是,文中假设疾病的治疗药物是未知的,因此所有药物(化学分子)都可以作为该疾病的候选药物. 另外,在构造集合 π^l 时,所有和候选药物 $drug_{\text{potential}}$ 可能有关的靶标都被当作该药物的候选靶标进行验证.

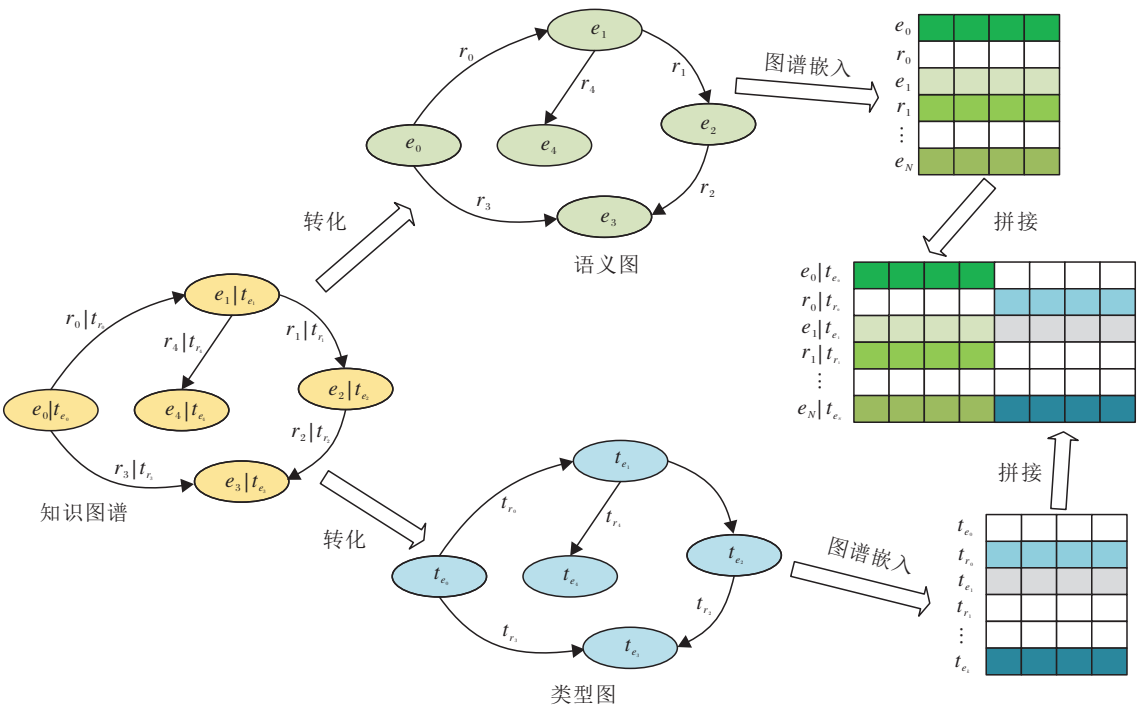


图 3 基于知识图谱分解与嵌入的向量转化

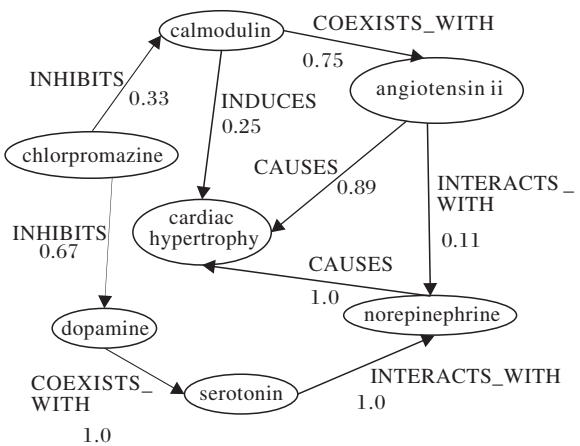
Fig. 3 Vector transformation based on knowledge graph splitting and embedding

1.4 药物发现的对比方法

本文采用部分机器学习方法和基于随机游走(Random Walk Algorithm, RWA)的药物发现方法作为对比方法. 机器学习方法包括逻辑回归(Logistic Regression, LR)、随机森林(Random Forest, RF)[9]和支持向量机(Support Vector Machine, SVM)[10]. 上述方法均有学者在药物发现实验中使用过. 随机游走的方法包括基本的 RWA 和两个目前效果最好的基于 RWA 的药物重定位方法(基于网络的异构网络重启随机游走(Network-Based Random Walk with Restart on the Heterogeneous Network, NRWRH)[11]和异构网络双向重启随机游走(Two-Pass Random Walk with Restart on a Heterogenous Network, TP-NRWRH)[12]).

使用基于 RWA 的药物发现方法发现疾病 $disease_i$ 的候选药物的过程为:首先,RWA 的起始节点设置为候选药物 $drug_{potential}$,RWA 的步数设置为 n. 然后,$drug_{potential}$-$disease_i$ 关系的分数可以被相应的随机游走算法给出. 最终,所有的候选药物通过对应分数给出排序.

图4为基于RWA发现氯丙嗪(Chlorpromazine)治疗心肌肥厚(Cardiac Hypertrophy)的过程,(a)为一个包含 7 个节点 9 条边的语义图,(b)为以 chlorpromazine 为起始点的RWA的结果. 如图所示,当 RWA 的步数为 1 时,chlorpromazine 无法走到节点 cardiac hypertrophy,因此 chlorpromazine 在 step_1 RWA(步数为 1 的 RWA)中分数为 0. 相似地,在 step_2 RWA 中 chlorpromazine-cardiac hypertrophy 关系的分数为 0.082 5. 当 RWA 的步数为 4 时,chlorpromazine 作为治疗 cardiac hypertrophy 的候选药物分数为 0.697.



(a)语义图

(a)Semantic Graph

	angiotensin ii	cardiac hypertrophy	calmodulin	chlorpromazine	dopamine	norepinephrine	serotonin
step_1	0	0	0.33	0	0.67	0	0
step_2	0.2475	0.0825	0	0	0	0	0.67
step_3	0	0.2202	0	0	0	0.697	0
step_4	0	0.697	0	0	0	0	0

(b)RWA 结果

(b)Results of RWA

图 4　基于随机游走的药物发现方法示例

Fig. 4　Illustration of drug discovery process by RWA-based method

2　实验及结果分析

2.1　实验数据

本文使用 PubMed 中 2013 年 6 月以前的所有生物医学摘要构造 SemKG 的数据集. 为了保证 SemKG 的质量,只被抽取一次的实体关系不用于构造该知识图谱. 表 1 为 SemKG 及其相关属性.

表 1　生物医学知识图谱 SemKG 属性表

Table 1　The detailed information of biomedical knowledge graph SemKG

SemKG 的数据	数量
PubMed 摘要	22769789
抽取出的实体关系	39133975
过滤后的实体关系	17651279
SemKG 中的实体	1067092
SemKG 中的关系	14419744
实体类型	133
关系类型	52

实验中的所有标准药物疾病关系都来自 Therapeutic Target Database(TTD)数据库. 首先使用 558 个 drug-target-disease 三元组作为标准药物疾病关系构造 93 230 个正例(路径长度 l 设为 4)[13]. 再通过随机替换标准药物疾病关系中的实体构造 drug′-target′-disease′ 三元组,进而构造相同数量的负例数据.

2.2　十倍交叉验证

本文采用十倍交叉验证评价方法性能,如表 2 所示. 在表 2 中,相比 LR、RF 和 SVM,本文采用的基于时间序列的方法能更好地预测疾病的候选药物.

表 2　不同方法的性能对比

Table 2　Performance comparison of different models

方法	准确率	召回率	F 值
LR	0.784	0.803	0.793
RF	0.753	0.864	0.804
SVM	0.669	0.857	0.751
BLSTM	0.930	0.948	0.939

为了验证知识图谱嵌入对本文方法的影响,表 3 对比使用不同嵌入层 BLSTM 模型的效果,包括随机嵌入($BLSTM_{random}$)、语义图嵌入($BLSTM_{SG}$)和类型图嵌入($BLSTM_{TG}$). 在 $BLSTM_{random}$ 中,实体和关系使用随机向量表示. $BLSTM_{TG}$ 和 $BLSTM_{SG}$ 分别使用类型图和语义图的图嵌入表示实体和关系.

由表 3 可知,尽管 $BLSTM_{TG}$ 只使用类型图的嵌入向量,仍取得一定效果. 这是因为语义图嵌入可以学习实体类型和关系类型间的规则. 然而,$BLSTM_{TG}$ 无法区分具有相同语义类型的不同实体间的关系,$BLSTM_{SG}$ 对性能具有较大提高. 基于知识图谱嵌入的方法 $BLSTM_{SG+TG}$ 取得最好结果,这是因为对知识图谱的嵌入既考虑图的结构信息,也考虑实体的语义信息.

表 3 基于不同图嵌入层模型的结果对比

Table 3 Result comparison of models with different graph embedding layers

方法	准确率	召回率	F 值
$BLSTM_{random}$	0.599	0.732	0.659
$BLSTM_{TG}$	0.783	0.859	0.819
$BLSTM_{SG}$	0.878	0.902	0.889
$BLSTM_{SG+TG}$	**0.930**	**0.948**	**0.939**

模型中的超参数设置如表 4 所示. 本文实验模型采用 adam 优化方法(adam Optimizer)优化,学习衰减率(Learning Rate Decay)设为 0.99,批量数据(Mini Batch)设为 1 000.

表 4 实验中的超参数选择

Table 4 Hyper-parameters selection in experiments

方法	范围	步数	最优值
BLSTM 细胞维度	[25,150]	25	75
BLSTM dropout	[0.3,0.8]	0.05	0.5
隐藏层维度	[10,100]	10	50
L_{SG}	[25,100]	25	100
L_{OG}	[10,50]	10	30

2.3 药物发现实验

本次实验从 TTD 中选择 115 例 drug-disease 关系. 每例数据表示已知该药物治疗该疾病,但药物作用机制尚不明确. 针对每种疾病 $disease_i$,随机从 TTD 中选择 100 种药物或化学物质作为该疾病的候选药物. 对每种疾病的 101 种候选药物(已知标准治疗药物和另外 100 种随机选择的候选药物)打分,其中标准答案 $drug_i$ 的平均排名(Meaning Ranking)和 115 例药物疾病关系中排名前 10 的百分比(Hits@10)作为模型的性能指标. 对于特定的疾病 $disease_i$,如果未找到标准治疗药物 $drug_i$,则该药物对应的分数为 0,排名为 101.

实验中 RWA 的步数设为 1 ~ 6,NRWRH 和 TP-NRWRH 的参数按原文中设置. 特别地,对于 GrEDeL,由于每种药物的作用机制尚不明确,从 TTD 中选择 5 785 个靶标(蛋白质、肽或核酸)作为每种药物的候选靶标. 因此,针对每种候选药物 $drug_i$ 和 $disease_i$,GrEDeL 构造 5 785 个 $drug_x$-$target_{potential}$-$disease_i$ 进行打分,最高分数作为该候选药物的分数. 表 5 为该药物发现实验的结果.

表 5 药物发现实验结果

Table 5 Results of drug discovery experiment

方法	未找到标准药物数量	平均排名	Hits@10/%
RWA_1	93	90.82	17.39
RWA_2	52	44.31	24.46
RWA_3	4	30.33	23.48
RWA_4	0	33.84	18.26
RWA_5	0	35.27	14.78
RWA_6	0	39.14	11.30
NRWRH	26	30.17	26.09
TP-NRWRH	19	29.87	27.83
BLSTM	0	**29.09**	**43.09**

由表 5 可知,RWA_1 只找到 22(115 - 93)个标准药物,这表明在知识图谱 SemKG 中只有 22 个药物和疾病直接相连. 当 RWA 的步数增加时,找到更多的标准药物,当步数超过 3 时,RWA 可以找到所有标准药物. 该结果表明所有标准药物和疾病在 SemKG 中都可以通过长度大于 3 的路径相连. NRWRH 和 TP-NRWRH 分别有 26 和 19 个标准药物没有找到. 虽然 NRWRH 和 TP-NRWRH 的步数都为 3,但由于其是基于重启的随机游走方法,重启可能会导致标准药物在 3 步内无法到达疾病节点.

对于平均排名指标,RWA_1 取得最差结果(90.82),这是因为该方法有 93 个标准药物无法找到. 当步数增加到 2 时,平均排名降低到 44.31. 这是因为 RWA_2 比 RWA_1 找到更多的标准药物. 但当步数继续增加时,平均排名指标也随着增大. 这是因为尽管步数超过 3 时所有的标准药物都可以找到,但同时更多错误的候选药物也会被发现. 对同一种疾病,更多被发现的候选药物可能会降低标准药物的排名. NRWRH 和 TP-NRWRH 的结果优于 RWA,这是因为 NRWRH 和 TP-NRWRH 根据生物医学知识选择随机游走的下一步,而不是完全随机的选择.

对于 Hits@10 指标,随着 RWA 步数的增加,Hits@10 值降低,说明 RWA 步数越多准确率越低. 相比 RWA,NRWRH 和 TP-NRWRH 取得更好的效果.

本文方法在各项指标中都取得最好的结果.

GrEDeL 能找到所有药物,具有较高的准确率(平均排名为 29.09,Hits@10 为 43.09%). GrEDeL 优于其它方法的原因在于:1)GrEDeL 覆盖所有长度为 2 ~ 4 的路径. 2)GrEDeL 结合知识图谱的结构信息和实体的语义信息预测药物疾病关系.

2.4　示例

表 6 为 4 个 GrEDeL 发现的药物及其作用机制. TTD 数据库已经报告碘克酸(Ioxaglate)可以治疗冠心病(Cardiovascular Disease),但其作用机制尚不明确. 本文方法从文献中找到 ioxaglate 是最有可能治疗冠心病的药物,并提供其作用机制为 ioxaglate 通过破坏血小板聚集进而影响冠心病中的信号转导途径,最终作用于冠心病. GrEDeL 发现依立曲坦(Eletriptan)可以治疗偏头痛(Migraine),其作用机制为 eletriptan 通过影响 trpa1 有关的感觉神经元进而治疗偏头疼. 由于这些药物的作用机制都是未知的,本文只能通过 GrEDeL 抽取的实体间关系对相关机制做出解释.

表 6　药物发现示例

Table 6　Examples for drug discovery

疾病	药物	排序	分数	作用机制
cardiovascular disease	ioxaglate	1	0.57	ioxaglate DISRUPTS platelet aggregation AFFECTS signal transduction pathway AFFECTS cardiovascular disease
migraine	eletriptan	6	0.55	eletriptan DISRUPTS sensory neuron PART_OF trpa1 ASSOCIATED_WITH migraine
Parkinson's disease	pramipexole	5	0.69	pramipexole AUGMENTS muscle PART_OF flap LOCATION_OF fall ISA parkinson's disease
dementia	rx-77368	9	0.51	rx-77368 NEG_AFFECTS blood flow AFFECTS hsf1 AFFECTS disease COEXISTS_WITH dementia

3　结束语

本文提出融合知识图谱和深度学习的文献挖掘药物发现方法. 相比现有方法,本文方法可以更有效地从文献中挖掘疾病的候选药物,提供相应的药物作用机制. 下一步的工作将通过不同的关系抽取技术构造更准确的生物医学知识图谱. 同时,提出图嵌入方法及基于 RNN 的深度学习方法,提高实验效果.

参考文献

[1] MORGAN S, GROOTENDORST P, LEXCHIN J, *et al*. The Cost of Drug Development: A Systematic Review. Health Policy, 2011, 100(1): 4-17.

[2] SPANGLER S, WILKINS A D, BACHMAN B J, *et al*. Automated Hypothesis Generation Based on Mining Scientific Literature // Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2014: 1877-1886.

[3] COHEN T, WIDDOWS D, SCHVANEVELDT R W, *et al*. Discovery at a Distance: Farther Journeys in Predication Space // Proc of the IEEE Conference on Bioinformatics and Biomedicine Workshops. Washington, USA: IEEE, 2012: 218-225.

[4] AHLERS C B, HRISTOVSKI D, KILICOGLU H, *et al*. Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms [C/OL]. [2018-09-25]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655783/pdf/amia-0006-s2007.pdf.

[5] HRISTOVSKI D, FRIEDMAN C, RINDFLESCH T C, *et al*. Exploiting Semantic Relations for Literature-Based Discovery[C/OL]. [2018-09-25]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839258/pdf/AMIA2006_0349.pdf.

[6] GAWEHN E, HISS J A, SCHNEIDER G. Deep Learning in Drug Discovery. Molecular Informatics, 2016, 35(1): 3-14.

[7] RINDFLESCH T C, FISZMAN M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. Journal of Biomedical Informatics, 2003, 36(6): 462-477.

[8] BORDES A, USUNIER N, GARCIA-DURAN A, *et al*. Translating Embeddings for Modeling Multi-relational Data // BURGES C J C, BOTTOU L, WELLING M, *et al*., eds. Advances in Neural Information Processing Systems 26. Cambridge, USA: The MIT Press, 2013: 2787-2795.

[9] CAO D S, ZHANG L X, TAN G S, *et al*. Computational Prediction of Drug Target Interactions Using Chemical, Biological, and Network Features. Molecular informatics, 2014, 33(10): 669-681.

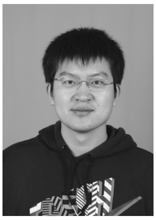
[10] BYVATOV E, FECHNER U, SADOWSKI J, *et al*. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. Journal of Chemical Information and Computer Sciences, 2003, 43(6): 1882-1889.

[11] CHEN X, LIU M X, YAN G Y. Drug-Target Interaction Prediction by Random Walk on the Heterogeneous Network. Molecular BioSystems, 2012, 8(7): 1970-1978.

[12] LIU H, SONG Y L, GUAN J H, *et al*. Inferring New Indications for Approved Drugs via Random Walk on Drug-Disease Heterogenous Networks[C/OL]. [2018-09-25]. https://doi.org/10.1186/s12859-016-1336-7.

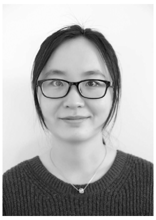
[13] SANG S, YANG Z, WANG L, *et al*. Sematyp: A Knowledge Graph Based Literature Mining Method for Drug Discovery[C/OL]. [2018-09-25]. https://doi.org/10.1186/s12859-018-2167-5.

作者简介



桑盛田,博士研究生,主要研究方向为数据挖掘、隐含知识发现. E-mail:sangst@ mail. dlut. edu. cn.
(**SANG Shengtian**, Ph. D. candidate. His research interests include data mining and literature-based discovery.)

杨志豪(通讯作者),博士,教授,主要研究方向为自然语言处理、数据挖掘. E-mail:yangzh@ dlut. edu. cn.
(**YANG Zhihao** (Corresponding author), Ph. D., professor. His research interests include natural language processing and data mining.)



刘晓霞,博士研究生,主要研究方向为数据挖掘、自然语言处理. E-mail:liuxiaoxia@ mail. dlut. edu. cn.
(**LIU Xiaoxia**, Ph. D. candidate. Her research interests include data mining and natural language processing.)

王 磊,博士,教授,主要研究方向为生物信息学. E-mail:wangleibihami@ gmail. com.
(**WANG Lei**, Ph. D., professor. Her research interests include biomedical informatics.)

赵 迪,博士研究生,主要研究方向为数据挖掘、自然语言处理. E-mail:zhao_di@ mail. dlut. edu. cn.
(**ZHAO Di**, Ph. D. candidate. His research interests include data mining and natural language processing.)

林鸿飞,博士,教授,主要研究方向为社交媒体数据挖掘、人工智能. E-mail:hflin@ dlut. edu. cn.
(**LIN Hongfei**, Ph. D., professor. His research interests include social media data mining and artificial intelligence.)

王 健,博士,教授,主要研究方向为自然语言处理. E-mail:wangjian@ dlut. edu. cn.
(**WANG Jian**, Ph. D., professor. Her research interests include natural language processing.)

* *

(上接 1095 页)

本届中国自动化大会还特别设立了大会研讨会,由大会组织委员会主席、西安交通大学管晓宏院士主持,航天科技集团吴宏鑫院士、东北大学柴天佑院士、中南大学桂卫华院士、华东理工大学钱锋院士、同济大学陈杰院士、北京交通大学宁滨院士、中国空间技术研究院杨孟飞院士,8 位院士面对面就“未来自动化”展开精彩讨论,并与参会代表互动。

会议第二天,中国航天科技集团公司科技委主任、中国科学院院士包为民作题为《发展航天智能技术,走进太空经济时代》的报告。中国科学院院士、“国家杰出青年科学基金”获得者、教育部“长江学者特聘教授”房建成院士妙语连珠,为大会带来第二个报告《原子陀螺仪技术发展与展望》。香港科技大学讲座教授王煜的报告题目是《智能无人技术与工业应用》。最后一个“压轴”报告来自中国自动化学会理事长、西安交通大学人工智能与机器人研究所教授、IEEE Fellow、中国工程院院士郑南宁,题目为《人工智能的基础研究做什么——从人类的大脑寻求人工智能发展的灵感》。

(下转 1119 页)