基于 SGCN 的化合物致癌性预测模型*

魏若冰，何家峰，邱晓芳，刘　旗
（广东工业大学 信息工程学院，广东 广州 510006）

摘　要： 癌症患者的激增引起了全世界的关注，许多研究者将目光放在了对化合物致癌性的评估上，但这是一项极其具有挑战性的任务。本实验获取了 341 种实验数据，利用三维图卷积网络(SGCN)，建立了对化合物致癌性的预测模型。结果表明：对化合物进行致癌性预测的 SGCN 分类模型准确率高达 96.9%，比其余模型效果更好，这表明 SGCN 模型能够准确地对化学品进行分类，并且在实际应用中具有相当大的潜力。

关键词： 三维图卷积网络；分类模型；致癌化合物
中图分类号： TP183　　**文献标识码：** A　　DOI：10.16157/j.issn.0258-7998.212080

中文引用格式： 魏若冰，何家峰，邱晓芳，等. 基于 SGCN 的化合物致癌性预测模型[J].电子技术应用，2022，48(6)：33-35，41.
英文引用格式： Wei Ruobing，He Jiafeng，Qiu Xiaofang，et al. Predict the carcinogenicity of compounds with SGCN[J]. Application of Electronic Technique，2022，48(6)：33-35，41.

Predict the carcinogenicity of compounds with SGCN

Wei Ruobing，He Jiafeng，Qiu Xiaofang，Liu Qi
（College of Information Engineering，Guangdong University of Technology，Guangzhou 510006，China）

Abstract： The rapid increase of the number of cancer patients has attracted worldwide attention. Researchers are very concerned about the assessment of the carcinogenicity of compounds，but this is extremely challenging. In this paper，341 kinds of experimental data were obtained，and the spatial atom feature combined with the spatial graph convolutional network(SGCN) was used to establish a model that could predict the carcinogenicity of compounds. The results showed that when compared to other models，the classification model of the SGCN was more suited to predicting the carcinogenicity of compounds and had an overall classification accuracy of 96.9%，which showed that the SGCN model could accurately classify chemicals and had considerable potential in practical applications.

Key words： spatial graph convolutional network；classification model；carcinogenicity of compounds

0 引言

由于技术的发展，新化合物的合成速度加快，每年诞生的化合物数以万计[1-2]，传统的评价方法不可能对所有的化合物进行评估。并且近年来患癌人数不断增多[3]，目前仍不清楚大多数的癌症是由于暴露于何种致癌化合物而导致的。世界卫生组织国际癌症机构(IARC)致癌清单中只有 429 种化合物被归为具有致癌性物质，但仍有 500 余种化合物未进行判定。传统的化合物致癌性评估主要通过实验测试进行，试验周期长且成本昂贵，不确定因素过多，因此迫切需要开发替代方法和工具来评估化合物的致癌性。

利用计算机进行毒性预测[4]是安全评价的重要手段，能够大幅度节省非临床安全评价试验成本，提高试验设计的科学性和准确性。随着机器学习的不断发展，支持向量机(SVM)、随机森林、神经网络(Random Forest)和 K-最近邻(KNN)等机器学习算法已被广泛用于化合物毒性预测中[5-7]。此外，对致癌性化合物的预测也有一些报道。2004 年，张晓昀等人[8]用人工神经网络中误差反向传播网络(BPNN)和径向基函数网络(RBFNN)对化合物的致癌性强弱进行了分类，模型的分类准确率达到了 80%以上；2005 年，张振山等人[9]用 PCA 对分子描述符降维，利用决策森林的方法预测化合物致癌性；在 2007 年，谢莹等人[10]基于 gSpan 算法，挖掘与已知毒性化合物具有相同字结构的化合物，进行未知化合物的毒性预测；2017 年，梁倩倩等人[11]基于量化构效关系(QSAR)方法预测 N-亚硝基化学物(NOCs)的致癌性，同年，阎爱侠等人[12]构建化合物的多维描述符，分别采用 4 种机器学习方法(朴素贝叶斯、随机森林、多层感知机和支持向量机)，模型的平均正确率达到 74%±3%。

近年来，越来越多的研究人员把目光转向致癌化合物的研究，但是现有的模型评估化合物的致癌性能力有限。本研究从多个数据库整理了化合物致癌性数据，基

* 基金项目：国家自然科学基金(61571140)

于具有空间结构的原子特征建立了三维图卷积网络(Spatial Graph Convolutional Network,SGCN)。

1 数据和方法

1.1 数据收集

从世界卫生组织国际癌症机构(IARC)致癌清单和美国环境保护局(EPA)列出的安全化合物清单(SCIL)中收集数据。为了保证数据的准确性和可靠性,用以下标准来筛选和处理数据:(1)IARC 致癌清单中选择有足够证据证明对人类具有致癌性的化合物,剔除其他分类中对致癌证据有限和致癌证据不足的化合物;(2)SCIL 安全化合物清单中选择根据实验和建模数据,已被证实不具有致癌性的化合物;(3)从上述条件筛选的数据集中剔除无法确定分子结构的化合物。最终,获得了 341 种实验数据,其中 246 种致癌性数据为正样本,余下 95 种不具有致癌性的数据为负样本,形成了最终的数据集。

1.2 数据集划分

从正负数据集中随机抽取数据:80%作为训练数据集(273 个分子)用于训练模型,10%作为验证数据集(34 个分子)用于调整超参数,10%作为测试数据集(34 个分子)用于评估模型的性能。

1.3 分子编码

采用独立热(one-hot)对原子特征进行编码[13]。独热编码又称一位有效编码,其方法是使用 N 位状态寄存器来对 N 个状态进行编码,每个状态都有独立的寄存器位,并且在任意时候,其中只有一位有效,如图 1 所示。同时,用 RDKIT 计算原子和键的特征,包括原子的符号、原子连接的键的个数、原子的价态和键的类型、是不是共轭、在不在环中等。

1.4 SGCN

本文中将分子的空间特性与传统的 GCN 相结合,去预测分子的致癌性。大多 GCN 模型使用二维分子图作为输入,通过特征矩阵和邻接矩阵去预测分子的性质[14]。然而,分子性质很大程度上受到空间中原子间相对位置影响,因此,在构建 SGCN 模型时,把带有原子坐标的分子图也作为输入。

$$z_{s\to s,i}^{(l)}=\mathrm{ReLU}[W_{s\to s}(s_i^{(l)}\|s_j^{(l)})+b_{s\to s}] \tag{1}$$

$$z_{v\to s,i}^{(l)}=\mathrm{ReLU}[(W_{v\to s}(\boldsymbol{V}_i^{(l)}\|\boldsymbol{V}_j^{(l)})+b_{v\to s})\cdot \boldsymbol{r}_{ij}] \tag{2}$$

$$\boldsymbol{Z}_{v\to v,i}^{(l)}=\tanh[W_{v\to v}(\boldsymbol{V}_i^{(l)}\|\boldsymbol{V}_j^{(l)})+b_{v\to v}] \tag{3}$$

$$\boldsymbol{Z}_{s\to v,i}^{(l)}=\tanh[(W_{s\to v}(s_i^{(l)}\|s_j^{(l)})+b_{s\to v})\otimes \boldsymbol{r}_{ij}] \tag{4}$$

式中,$s_i^{(l)}$、$s_j^{(l)}$和 $\boldsymbol{V}_i^{(l)}$、$\boldsymbol{V}_j^{(l)}$分别表示在 l 层上第 i、j 个原子的标量特征和矢量特征,W 和 b 是每一次迭代的权重和偏置,$\boldsymbol{r}_{ij}$ 是原子 i、j 间的相对位置矩阵,z 和 Z 为更新后的标量和矢量;$\|$是将两个特征连接起来,$\cdot$表示点积,$\otimes$表示张量积。一旦对标量和矢量特征进行了变换,4 个更新后的特征首先以矢量加矢量、标量加标量的形式拼接起来,然后将它们线性组合在一起。

$$s_i^{(l+1)}=\mathrm{ReLU}[\sum_{j\in\{i,(N(i))\}}\boldsymbol{A}_{ij}(W_s(z_{s\to s,i}^{(l)}\|z_{v\to s,i}^{(l)})+b_s)] \tag{5}$$

$$\boldsymbol{V}_i^{(l+1)}=\tanh[\sum_{j\in\{i,(N(i))\}}\boldsymbol{A}_{ij}(W_v(z_{v\to v,i}^{(l)}\|z_{s\to v,i}^{(l)})+b_v)] \tag{6}$$

式中,$\boldsymbol{A}$ 是标准化后的邻接矩阵,W 和 b 表示权重和偏置。空间 GCN 由卷积层、特征构造层和全连接层 3 个模块组成,如图 2 所示。

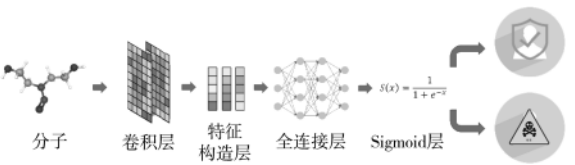

图 2 空间图卷积流程图

在初始化特征时,节点的标量特征进行独立热编码形成 60 个特征,而矢量特征被初始化为零。卷积层的第一阶段将每个节点的两个特征融合在一起,生成中间特征。在第二阶段,收集中间特征并沿着邻域进行汇总,从而产生更高级别的特征。通过卷积层,更新标量特征和矢量特征。经过卷积后,特征构造层通过两种策略收集节点上的特征:$\mathrm{SGCN_{sum}}$ 整合了节点上分布的所有原子特征,生成分子的标量和矢量特征;$\mathrm{SGCN_{max}}$ 选取原子特征中取值最大的作为分子特征。生成的分子特征被送到具有 ReLU 激活的全连接神经网络。最后,输出被扁平化处理后送到单层神经网络中来进行分类。

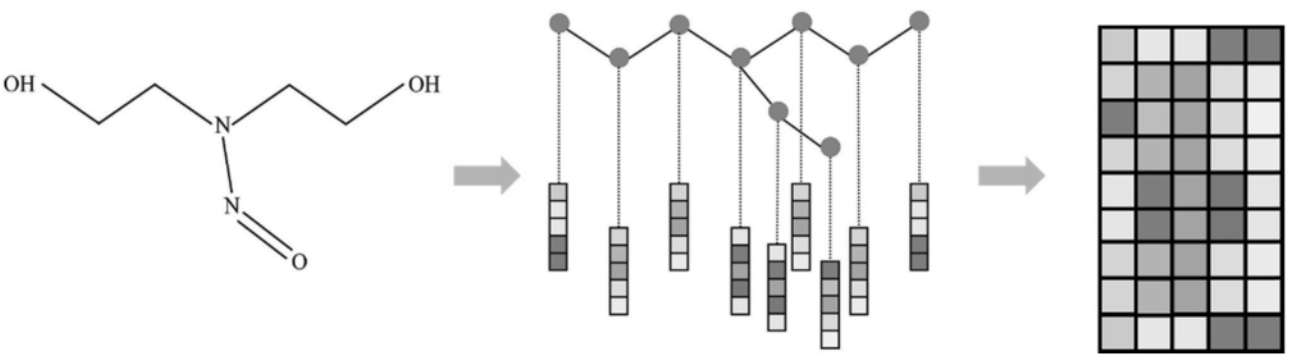

图 1 对分子图进行独立热编码示意图

1.5 对比模型

对比模型包括GCN、多层感知机(Multilayer Perceptron,MLP)、随机森林(Random Forest,RF)、支持向量机(Support Vector Machines,SVM)、K-最近邻算法(K-Nearest Neighbors,KNN)、决策树(Decision Tree)、线性判别分析(Linear Discriminant Analysis,LDA)和XGBoost。GCN模型由两个卷积层和一个全连接层构成，学习率为0.001。多层感知机中设置优化权重设置为adam，最大迭代300次。余下机器学习模型从scikit-learn库中调用，随机森林中建立子树的数量为20；支持向量机中核函数类型为径向核函数，布尔值为Truth；朴素贝叶斯分类器中拉普拉斯平滑系数设置为1，其余模型参数均设置为默认值。

1.6 模型评估方法

采用10折交叉验证法来评估模型的预测性能和可靠性。在10折交叉验证中，先将数据集划分为10个大小相等的互斥子集，每个子集都尽可能保持了数据分布的一致性，之后，每次都用9个子集作为训练集，余下的1个子集作为验证集。然后，将交叉验证过程重复10次。

$$Acc=(TP+TN)/(TP+TN+FP+FN) \quad (7)$$

$$Pre=FP/(FP+TN) \quad (8)$$

式中，TP是真阳性，TN是真阴性，FP是假阳性，FN是假阴性。计算总体预测准确率(Acc)以对每个预测函数进行评估。此外，为了使评价更有效，加入了查准率(Pre)来进一步验证模型。查准率是计算模型判断为阳性的样本中有多少是真正的阳性。

2 结果和讨论

2.1 空间GCN特征构造

在特征构造层以两种方式对特征进行构造，一种是将分布在节点上的所有原子特征相加($SGCN_{sum}$)，形成新的矢量和标量特征；另一种是选取最大值的原子特征作为分子特征($SGCN_{max}$)，依据范数比对矢量特征的大小。根据表1可以看出，$SGCN_{max}$和$SGCN_{sum}$在对模型准确率预测在0.946~0.973之间，查准率在0.939~0.951之间，在GCN为基础上准确率和查准率提高了约4.5%。在特征构造上，对比模型$SGCN_{max}$和$SGCN_{sum}$在评估参数上的值，可以发现，$SGCN_{sum}$除了在验证集的准确率略微低于$SGCN_{max}$，其余均高于$SGCN_{max}$，所以，在对分子致癌性进行预测时，特征构造中选取原子特征的最大值会使得模型效果偏好。

表1 模型的评估指标

模型	Acc		Pre	
	验证集	测试集	验证集	测试集
$SGCN_{max}$	0.973	0.969	0.951	0.944
$SGCN_{sum}$	0.942	0.946	0.944	0.939
GCN	0.912	0.905	0.928	0.921

2.2 对比实验

此外，还构建了7个预测模型作为对比，7个模型的整体准确率在0.810~0.861之间，如表2所示。

表2 基于原子特征和分子描述符的对比模型

模型	原子特征				分子描述符			
	Acc		Pre		Acc		Pre	
	验证集	测试集	验证集	测试集	验证集	测试集	验证集	测试集
RF	0.829	0.844	0.823	0.834	0.920	0.920	0.915	0.923
SVM	0.810	0.805	0.799	0.813	0.813	0.820	0.821	0.817
KNN	0.836	0.84	0.845	0.855	0.821	0.833	0.839	0.840
DT	0.821	0.803	0.814	0.825	0.889	0.879	0.865	0.854
LDA	0.861	0.85	0.857	0.849	0.822	0.913	0.905	0.915
XGBoost	0.810	0.813	0.820	0.814	0.931	0.929	0.924	0.928
MLP	0.845	0.841	0.850	0.839	0.843	0.858	0.861	0.852

在准确率评估中，表现最好的是RF模型为0.844；在查准率评估中，表现最好的则是KNN算法为0.855。在验证集中预测性能最好的LDA算法在测试集中的表现同样优异，其总体预测准确率为0.861，查准率为0.849。除此之外，KNN和RF也表现出了较好的预测能力，KNN在验证集的查准率达到了0.855。对比分析表2中的模型可以看出，验证集和测试集中总体预测准确率和查准率基本相等，表明模型不存在过拟合的现象。将此表模型中表现最好的几个模型同SGCN进行对比，可以看出SGCN表现出了较为优异的性能。

2.3 提取分子描述符

在与7种模型的对比实验中发现，与SGCN和GCN进行对比时准确率差异过大达到了0.109，考虑到所有模型提取的特征为原子特征，SGCN中的输入仅包括原子的特征矩阵还包括原子间的邻接矩阵和相对位置矩阵，而在对比实验中输入仅为原子特征，输入信息量相对较少且不全面，以用分子的信息代替原子的信息作为对比模型的输入。分子描述符[15]通过量化部分结构和物理化学性质来表达化合物的化学特征。使用函数调用rdkit生成数据集中所有分子的描述符，生成的描述符包含分子指纹、相对分子质量和部分电荷等200维特征。将分子描述符作为输入用于7种对比模型中，发现准确率有明显的上升，整个模型的准确率在0.821~0.931之间，其中验证集中RF和XGBoost的准确率分别从0.829和0.810上升至了0.920和0.931，除此之外DT的准确率也上升了0.6，其他模型准确率没有变化或略微下降。

3 结论

本研究采用SGCN模型对化合物进行了致癌性预测，可因此减少因条件限制而导致的化合物致癌性评估不足。此模型对273种数据集和34种外部验证数据集进行毒性分类，在34种测试集中获得了96.9%的准确

(下转第41页)

34(9):1324-1329.
[5] 李利利.基于 Verilog HDL 的 SPI 协议可复用 IP 软核的设计与验证[D].兰州:兰州大学,2015.
[6] 孙永涓.基于 AHB-Lite 总线的高速 SPI 接口的设计与实现[D].沈阳:辽宁大学,2021.
[7] 田晓旭,徐庆阳,汤先拓,等.基于 UVM 的寄存器验证自动化方法[J].集成电路应用,2020,37(2):18-21.
[8] 刘斌.芯片验证漫游指南[M].北京:电子工业出版社,2018:393-394.
[9] 任传宝,崔建国,鲁迎春,等.应用直接编程接口技术提高片上系统的 UVM 验证重用性[J].微电子学与计算机,2021,38(6):20-26,32.
[10] 李世超.基于 UVM 的 MC-SOC 中可重用验证平台的设计与实现[D].成都:电子科技大学,2018.

(收稿日期:2021-11-15)

作者简介:
刘森态(1996-),男,硕士研究生,主要研究方向:数字芯片验证、集成电路设计。
庞宇(1978-),通信作者,男,博士,教授,主要研究方向:集成电路设计,E-mail:pangyu@cqupt.edu.cn。
魏东(1997-),男,硕士研究生,主要研究方向:数字芯片验证、集成电路设计。

(上接第 35 页)

率和 94.4%的查准率,表现出了评估化合物致癌性的优异性。通过进一步分析,发现用分子描述符作为特征时,RF 和 XGBoost 模型效果准确率也达到 90%以上,这两种模型同样也适用于化合物致癌性的分类。将 SGCN 模型用于有毒气体分类上,准确率达到 89%,说明此模型在化合物分类判定上也有一定的普适性。

该研究探索了基于原子空间特征结合 SGCN 构建化合物致癌性分类模型的可行性,为化学物的健康风险评估提供依据,然而收集到的样本数和样本类别有限,需进一步增加样本量,使构建出的模型具有更好的泛化性和稳定性。

参考文献

[1] 方从兵,宛晓春,江昌俊.黄酮类化合物生物合成的研究进展(综述)[J].安徽农业大学学报,2005,32(4):498-504.
[2] AMRI N,WIRTH T.Recent advances in the electrochemical synthesis of organosulfur compounds[J].The Chemical Record,2021,21(9):1-13.
[3] FERLAY J,COLOMBET M,SOERJOMATARAM I,et al.Cancer statistics for the year 2020:An overview[J].International Journal of Cancer,2021,149(8):1-12.
[4] 程飞雄,沈杰,李卫华,等.有机化合物的陆地和水生环境毒性的计算机预测研究[J].农药学学报,2010,12(4):477-488.
[5] KINGMA D,BA J.Adam:a method for stochastic optimization[C]//Computer Science,2014.
[6] JIANG C,YANG H,DI P,et al.In silico prediction of chemical reproductive toxicity using machine learning[J].Journal of Applied Toxicology,2019,39(6):844-854.
[7] TOROPOVA A P,TOROPOV A A,MARZO M,et al.The application of new HARD-descriptor available from the CORAL software to building up NOAEL models[J].Food & Chemical Toxicology An International Journal Published for the British Industrial Biological Research,2018,112:544-550.
[8] 张晓昀,马卫平,刘满仓,等.人工神经网络用于多环芳烃及胆蒽系化合物致癌活性的研究[J].兰州大学学报(自然科学版),2004,40(1):38-44.
[9] 张振山,罗小民,郑明月,等.采用基于决策森林的分类方法预测化合物致癌毒性[C]//中国化学会第九届全国量子化学学术会议暨庆祝徐光宪教授从教六十年论文摘要集,2005.
[10] 谢莹,吴建国,李炜,等.基于 gSpan 算法的未知化合物毒性预测[J].合肥工业大学学报(自然科学版),2007,30(10):1278-1280.
[11] 梁倩倩,郑唯韡,何更生,等.基于分类和交叉参照的改良量化构效关系预测 N-亚硝基化学物的致癌性[J].中华预防医学杂志,2017,51(7):621-627.
[12] 阎爱侠,孔越.基于多领域数据信息融合的化合物致癌性预测[C]//中国化学会第 14 届全国计算(机)化学学术会议暨分子模拟国际论坛,2017.
[13] XU Y,DAI Z,CHEN F,et al.Deep learning for drug-induced liver injury[J].Journal of Chemical Information and Modeling,2015,55(10):2085-2093.
[14] DEFFERRARD M,BRESSON X,VANDERGHEYNST P.Convolutional neural networks on graphs with fast localized spectral filtering[C]//Lausanne:EPFL,2016.
[15] 任伟,孔德信.定量构效关系研究中分子描述符的相关性[J].计算机与应用化学,2009,26(11):1455-1458.

(收稿日期:2021-08-20)

作者简介:
魏若冰(1996-),女,硕士研究生,主要研究方向:图卷积网络算法。
何家峰(1970-),通信作者,男,副教授,主要研究方向:图像处理与模式识别,E-mail:jfhe@gdut.edu.cn。
邱晓芳(1996-),女,硕士研究生,主要研究方向:卷积神经网络、仿生嗅觉。

版权声明