基于递归神经网络的药物分子ADMET性质预测

Report on 2019.12.14

**潘高翔，蔡丹杨，付思杰**

**北京大学化学与分子工程学院，北京大学工学院**

# 方法原理

## 2-1 输入数据结构

首先利用rdkit包提供的smiles解析器，将以smiles表达式格式存储的分子结构转化为分子图Graph数据结构(自定义无向图)，Graph由节点(Node)构成，Node对象对应原子，原子类型以one-hot编码的形式封装在Node中。Node对象还存储了与之连接的其他Node对象引用，即“边”，对应化学键，以及这些引用的权值，即化学键的类型，同样以one-hot编码的方式存储。

预测与训练中需应用分子树，可理解为具有根-枝-叶结构的有向无环分子图，同样由Node构成。Graph对象通过最小深度生成树算法(定义在Graph类下的build\_tree方法中)，可建立以任意Node为根节点的分子树，用于递归神经网络模型的预测和训练。

## 2-2 递归神经网络

模型输入分子图，输出ADMET预测值。模型可分为性质预测和特征抽取两部分，即：

E为特征抽取函数，返回一个定长特征向量，为单隐层经典神经网络映射，返回预测值。

E函数为求一组递归网络返回值的加和，即：

为递归神经网络映射，为以原子k为根节点的分子树的根节点，为以原子k为根节点的分子树中，k原子节点的返回值向量。

计算当前节点所有子节点在映射下返回值向量，将这些向量拼接上键连信息向量后利用隐层循环采样，并将采样向量求和得到定长向量，与当前节点的原子信息向量拼接，通过单隐层神经网络映射得到返回值向量，数学描述如下：

注意到这是一个递归函数，当递归进行至分子树的叶节点时，由于叶节点无子节点，网络开始逐级向上返回向量值。与Alessandro Lusci等人开发出的递归神经网络原理略有不同：

由于神经网络算法常要求向量长度确定，然而分子树中节点的子节点数目不定，此递归神经网络利用零向量将剩余的特征空间补齐。本文对这一点进行了改进，利用类似于卷积神经网络的方式，对各个子节点循环抽取向量并最终加和，得到定长向量。由于各子节点共享权值，此方法大幅减少了数据维度，且此方法将各子节点放在相同地位上，更加符合实际，也更优雅。

# 模型评估

## 2-1 运行环境

Intel® Core™ i5-9300H @CPU 2.40GHz 2.40GHz

Anaconda3, python3.7.4

## 2-2 参数设置：

以下是一代参数

1. 网络(递归内层网络)输入23维(20维+3维键向量)，输出20维，激活函数Tanh；
2. 网络(递归外层网络)输入34维(20维内层输出+14维元素向量)，对于叶节点，输入仅14维，无内层输入，输出20维，激活函数LeakyReLU；
3. 网络(性质预测网络)，输入20维，隐层25维，输出1维，激活函数LeakyReLU，因执行拟合任务，输出无激活函数；
4. Loss(损失函数)，采用MSE损失函数；
5. Optimizer(优化器)，使用Adam算法，学习率0.001，；
6. 训练集共858组数据，测试集286组，占比0.2，训练进行8轮。使用类似自助法采样，训练数据随机。计算机无GPU，训练过程全部在CPU上进行。

## 2-3 评估结果

训练用时5 min 26.04 seconds，测试集上

## 2-4 讨论分析

此模型为测试模型，未经调参，且与网络均未设置隐层，输出网络采用LeakyReLU这样的利于训练但表现不好的激活函数，因此最终拟合精度不尽人意，我们将把与网络设置为单隐层网络，但预期训练时间也将相应增长。

# 参考文献

1. Alessandro Lusci; Gianluca Pollastri; and Pierre Baldi, Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules, *J. Chem. Inf. Model.* **2013**, 53, 1563−1575;
2. Feixiong Cheng; Weihua Li; Yadi Zhou; Jie Shen; Zengrui Wu; Guixia Liu; Philip W. Lee; and Yun Tang, admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties, *J. Chem. Inf. Model.* **2012**, 52, 3099−3105.