人工神经网络在材料基因组中的应用

1 问题背景

材料对于现代工业而言意义重大。然而，现代的材料学研究还是基于传统的试错法，不仅效率低下，同时会造成大量的资源浪费。2011 年 6 月 24 日，美国总统奥巴马宣布启动了材料基因组计划（Material Genome Initiative），该计划旨在通过高通量计算、计算机模拟等手段辅助材料设计，加速新材料的研发周期，降低其研发成本。我国也紧随其后，若干高校先后成立了材料基因工程中心，众多科技公司也相继开展了计算机技术与新材料融合的工作。随着人工智能的高速发展，以及材料学海量的数据积累，利用机器学习算法，完成材料学领域数据挖掘，获得材料背后的信息，已经成为材料研发的新趋势。

2 问题描述

图 1 是二维分子的结构图，通过 python 提供的相关库，rdkit，以及 mordred 可以很快捷的获得分子的指纹信息。本案例分析的主要目的是利用神经网络模型建立分子指纹和分子性质的定量构效关系模型。

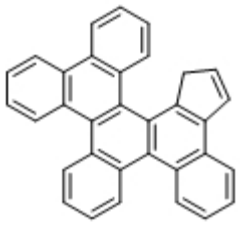

图 1 某分子的结构图

3 方法与结果

首先，根据需求选择化学分子描述符的种类，本案例提供 ECFP 以及 Mordred 两种分子描述符。这两种化学描述符在分子性质预测领域中表现良好，同时简单易得。之后，将分子描述符作为输入的特征用于机器学习模型的建立，通过对模型不断的优化，得到性能良好的分子性质预测模型。本案例中选取神经网络模型，神经网络的结构如图 2 所示

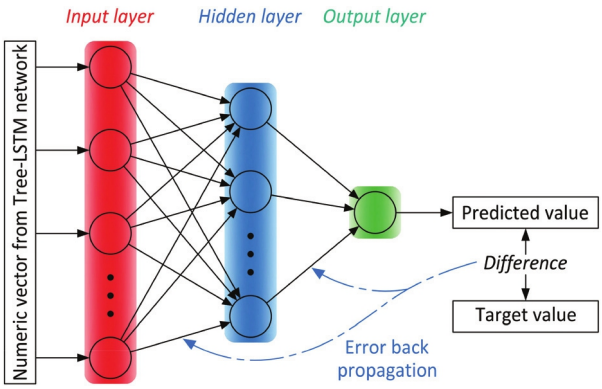

图 2 神经网络结构模型图

本案例准备的数据集为有机光伏材料分子的光电转化效率，共计 3000 个数据集。这里展示了部分的分子结构图

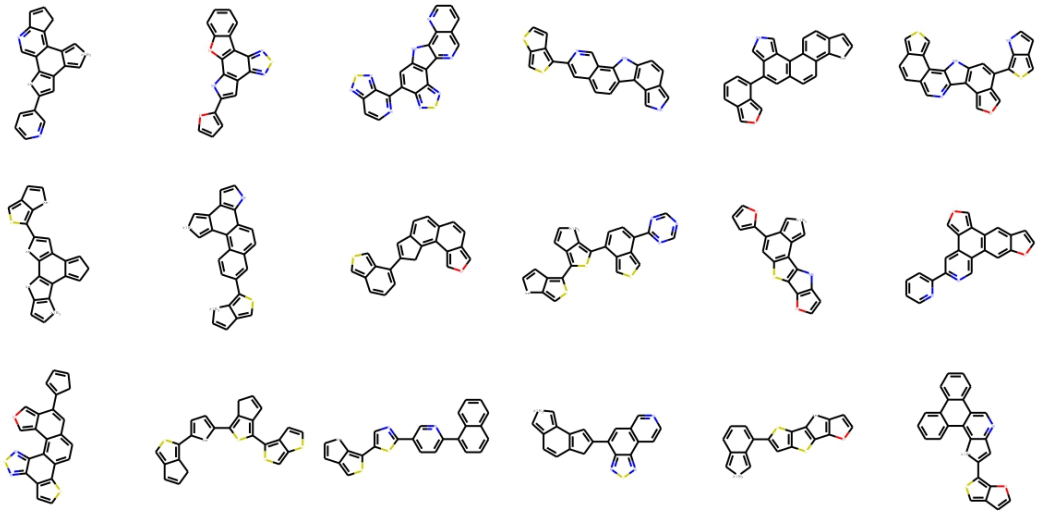

图 3 有机光伏材料结构（部分）

由于本案例的原始数据为分子的结构信息，是非结构化的信息，因此需要通过相关算法，将其转化为结构数据。这里利用 python 中两个第三方库，rdkit 以及 mordred，分别提取分子的 ECFP 描述符以及 mordred 描述符。其中 ECFP 描述符的维度为 512，Mordred 描述符的指纹为 489。分别将两种描述符作为神经网络的输入，并对神经网络进行训练和测试。这里选择 2400 个数据为训练数据，600 个数据为测试数据，通过网格搜索法对网络的超参数进行调优，最终确定隐含层数为 3，各隐含层的神经元数分别为 128，128，32，选取的激活函数为 relu 函数，优化器为 Adam 优化器

为了更好地验证模型的准确度，性能测试指标选择R^2、$RMSE$（均方根误差）和残差图，如方程(1)-(3)所示。

$$R^2 = 1 - \frac{\sum_{n=1}^{N}\left(y^{(n)} - f^{(n)}\right)^2}{\sum_{n=1}^{N}\left(y^{(n)} - \bar{y}\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}\left(y^{(n)} - f^{(n)}\right)^2}{N}} \tag{2}$$

$$\varepsilon^{(n)} = y^{(n)} - f^{(n)} \tag{3}$$

其中，$y^{(n)}$是x_n对应的输出值；$f^{(n)}$是x_n对应的预测值；$\bar{y}$输出值的平均值；$\varepsilon^{(n)}$是残差。

得到的验证结果图 4-5 所示。对于 ECFP 描述符而言，$RMSE = 0.432$，$R^2 = 0.936$，对于 Mordred 描述符而言，$RMSE = 0.5257$，$R^2 = 0.905$，通过两种描述符的对比可以发现，对于本案例，两种描述符都能有不错的预测效果，但是 ECFP 描述符的性能更优。

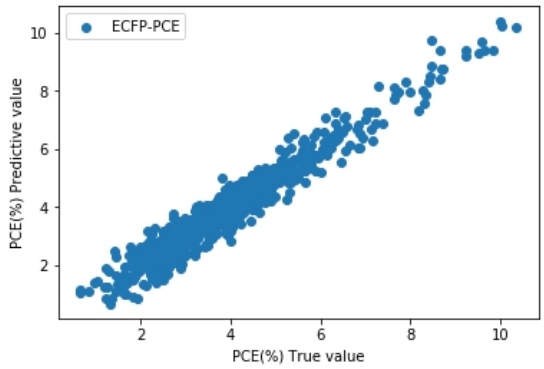

图 4 ECFP 描述符预测结果与真实结果对比图

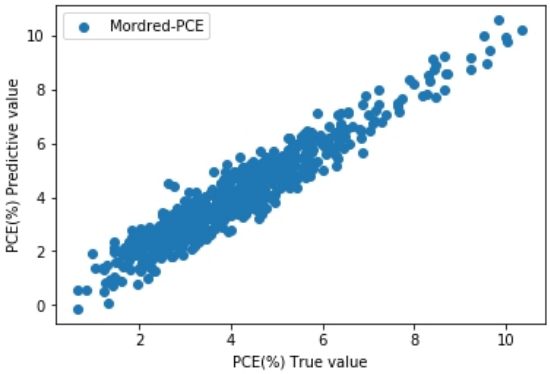

图 5 Mordred 描述符预测结果与真实结果对比图