|  |  |
| --- | --- |
|  | |
| **PyMolPredictor 项目设计文档**  **V1.0** | |
|  |

|  |
| --- |
|  |

**版 本 历 史**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本/状态 | 作者 | 参与者 | 日期 | 备注 |
| 1.0.0 | 李奡程、兰方舟、高睿齐 |  | 2020-1-11 | 创建 |

**目 录**

[第一部分 引言 6](#_Toc326322992)

[一、编写目的 6](#_Toc326322993)

[二、读者对象 6](#_Toc326322994)

[三、术语与缩写解释 6](#_Toc326322995)

[1、QSAR 6](#_Toc326322996)

[2、分子描述符 7](#_Toc326322997)

[2、SMILES 7](#_Toc326322997)

[2、RNN 7](#_Toc326322997)

[3、VAE 7](#_Toc326322997)

[第二部分 项目概述 11](#_Toc326323005)

[一、项目描述 11](#_Toc326323006)

[二、项目功能描述 11](#_Toc326323007)

[1、数据处理 12](#_Toc326323008)

[2、模型训练 12](#_Toc326323009)

[3、结果分析 13](#_Toc326323010)

[4、活性预测 13](#_Toc326323011)

[5、分子设计 14](#_Toc326323012)

[第三部分 设计约束 17](#_Toc326323021)

[一、需求约束 17](#_Toc326323022)

[1、用户界面标准 17](#_Toc326323026)

[2、软件质量 17](#_Toc326323027)

[二、隐含约束 19](#_Toc326323028)

[第四部分 前端设计 35](#_Toc326323059)

[一、前端整体结构 35](#_Toc326323060)

[二、实体描述 35](#_Toc326323061)

[1、Supplier实体描述 35](#_Toc326323062)

[2、SaleList实体描述 35](#_Toc326323063)

[第五部分 模型设计 35](#_Toc326323059)

[一、模型原理简述 35](#_Toc326323060)

[二、实体描述 35](#_Toc326323061)

[1、QSARDNN实体描述 35](#_Toc326323062)

[2、SmilesRNN实体描述 35](#_Toc326323063)

[第六部分 运行环境和部署 49](#_Toc326323101)

[一、运行环境 49](#_Toc326323102)

[二、系统性能要求 49](#_Toc326323106)

第一部分 引言

### 编写目的

|  |
| --- |
| 本文档编写目的是为读者提供PyMolPredictor软件的正确操作方法，浅显讲解用到的化学学科知识和深度学习技术，以期读者能够正确使用软件完成期望的功能，为正确运行和维护提供指引，同时允许用户根据自己的需求更改相关代码，完成更复杂的任务。 |

### 二、读者对象

|  |
| --- |
| 计算化学学科从业者，并对深度学习技术有一定的了解与兴趣；对深度学习有一定知识，想从事计算化学的计算机行业人员等。 |

### 三、术语与缩写解释

|  |
| --- |
| 1.QSAR: **定量构效关系（Quantitative Structure-Activity Relationship）**是一种借助分子的理化性质参数或结构参数，以数学和统计学手段定量研究有机分子生理相关性质的方法。这种方法广泛应用于药物、农药、化学毒剂等生物活性分子的合理设计。 |
| 2. 分子描述符：描述分子在某一方面性能的量，可作为分子特征向量的某维输入QSAR模型中建模。常用分子描述符有C原子个数、N原子个数、水溶性等。 |
| 3. SMILES： Simplified molecular input line entry specification，简化分子线性输入规范，是一种用ASCII字符串明确描述分子结构的规范。SMILES由Arthur Weininger和David Weininger于20世纪80年代晚期开发，现为计算化学常用分子表示形式之一。 |
| 4. RNN： 循环神经网络（Recurrent Neural Network, RNN）是一类以序列数据为输入，在序列的演进方向进行递归且所有节点按链式连接的神经网络，常用于文本处理等。在本软件中，用于SMILES串的处理和学习。 |
| 5. VAE：Variational Autoencoder的简称。变分自编码器作为自编码器的升级版本，除了通用的encoder和decoder结构外，其还通过规约向量在隐层空间（latent space）上符合高斯分布，使得学习的向量更加规范化，往往具有比单个encoder更好的性能。 |

第二部分 项目概述

### 一、项目描述

|  |
| --- |
| 本项目实现了一个具有图形界面的化学分子信息系统，该软件命名为PyMolPredictor，目前版本号是V1.0。 |

### 二、项目功能描述

|  |
| --- |
| 本项目能够完成以下五项功能： |
| 1. 数据处理：对于用户给定的包含QSAR关系的文件（默认格式为csv），进行训练集-测试集的划分，对数据中缺失值进行处理，形成较为干净的数据，可直接用于后续训练测试；对数据进行特征分析，并可视化数据在低维空间的分布，给用户以直观感受；对指定属性查看直方图分布，了解相应属性的分布等。 |
| 1. 模型训练：对于用户想要建模的QSAR，根据用户输入的模型参数进行训练并保存模型；支持用户在之前的模型的基础上继续训练，进行微调；可通过分子描述符向量进行建模（DNN），同时也可通过SMILES串进行建模（RNN）；为用户显示相关训练进度和细节等等。 |
| 1. 结果分析：对于用户的训练历史进行分析，根据用户选定的log可视化验证集上的损失函数变化曲线、验证集的预测结果分布和相关模型架构等。 |
| 1. 活性预测：根据用户指定的模型，在用户选定的数据上进行预测，以表格形式输出相关预测结果，并可视化相关活性最高的分子和预测结果分布等等。 |
| 1. 分子设计：根据用户指定的**包含**SMILES的数据，构建分子设计模型（训练时间较长，建议使用GPU机器）；使用指定的分子设计模型进行分子设计，并按活性高低可视化设计的结果。 |

第三部分 设计约束

### 一、需求约束

|  |
| --- |
|  |

### 二、隐含约束

|  |
| --- |
|  |

第四部分 前端设计

### 一、前端整体结构

|  |
| --- |
|  |

### 二、实体描述

|  |
| --- |
|  |

第五部分 模型设计

### 一、模型原理简述

|  |
| --- |
| QSAR深度学习模型使用多层感知机作为深度学习模型，可以进行回归预测与分类预测。回归模型使用L2 loss函数(均方误差)，分类模型使用CrossEntropy Loss函数(交叉熵)。模型为四层感知机，并且在中间层加入batchnorm批标准化，使得深度神经网络训练过程中每一层神经网络的输入保持相同分布的，加快模型收敛速度。模型激活函数为RELU函数，防止梯度下降现象。模型输入为分子描述符向量，在加载数据前需要先将数据集上的数据进行标准化，使得每一种属性在整个数据集上的分布变为均值为0，方差为1的标准化数据，否则过多离散值的引入会导致部分属性失去在模型中的作用，导致预测值完全相同。 |

### 二、实体描述

|  |  |
| --- | --- |
| 1.QSARDNN实体描述   |  | | --- | | 本项目中我们将QSAR深度学习模型封装为QSARDNN类，该类提供了train,test,load,save以及setPropertyNum五种接口。  setPropertyNum用于设定模型输入的维度(分子描述符的维度)以实例化多层感知机模型。train函数用于模型的训练,用户可以自己指定训练的学习率，批大小，训练的轮数以及是否需要earlystop；训练过程中程序会自动划分validation集合作为判断是否过拟合，若指定earlystop则当在validaion集合上预测效果下降时提前退出。  test函数使用当前训练好的模型进行预测,返回预测的标签集。  save和load函数分别负责保存和加载当前模型。 | |
|  |

第六部分 运行环境和部署

### 运行环境

|  |
| --- |
|  |

### 二、系统性能要求

|  |
| --- |
|  |