**《水环境遥感反演系统 V0.1.2》**

**使用说明书**

**目录**

[**第一章** **系统简介** 1](#_Toc10041157)

[**第二章** **系统安装说明** 2](#_Toc10041158)

[**1.安装要求** 2](#_Toc10041159)

[**1.1硬件环境** 2](#_Toc10041160)

[**1.2软件环境** 2](#_Toc10041161)

[**2.安装方法** 2](#_Toc10041162)

[**第三章** **系统使用说明** 3](#_Toc10041165)

[**1.栅格数据工具** 3](#_Toc10041166)

[**1.1像元值提取** 3](#_Toc10041167)

[**2.数据管理工具** 5](#_Toc10041168)

[**2.1创建样本** 5](#_Toc10041169)

[**2.2归一化/标准化** 6](#_Toc10041170)

[**2.3特征处理** 8](#_Toc10041171)

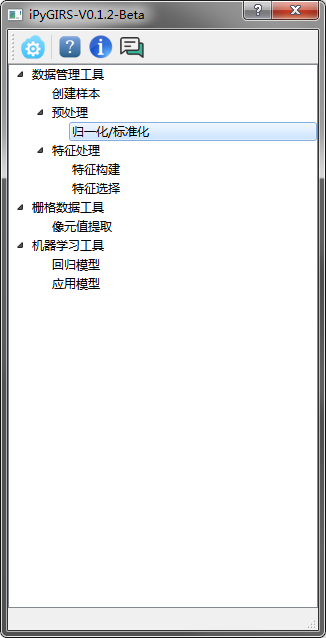
[**3.机器学习工具** 11](#_Toc10041172)

[**3.1回归模型** 11](#_Toc10041173)

[**3.2应用模型** 15](#_Toc10041174)

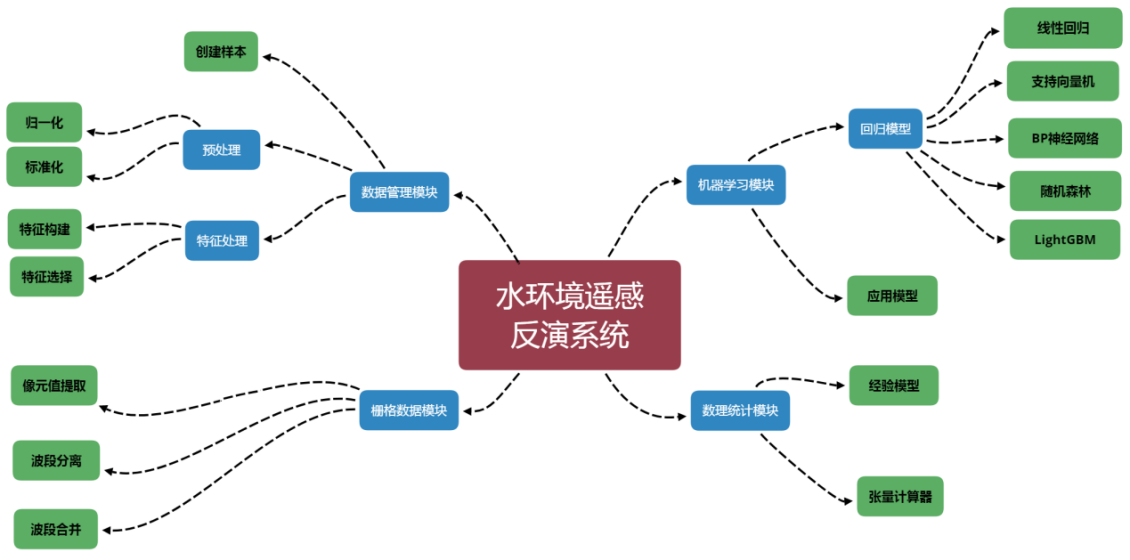
**第一章 系统简介**

水环境遥感反演系统（iPyGIRS）是一个具有数据批量化处理、模型建立与应用全自动化特点的软件。该系统以机器学习回归模型的训练测试为核心功能，集成了数据预处理、样本创建、遥感影像像元值提取、经验模型建模、水质参数预测等功能，可应用于多种水环境遥感监测场景中。



**图1-1-1 水环境遥感反演系统主界面**

程序功能模块如图1-1-2所示，当前版本实现了除除“数理统计”模块之外其它3个模块的绝大多数功能。



**图1-1-2 水环境遥感反演系统模块**

1. **系统安装说明**

《水环境遥感反演系统》完全采用Python开发，由于Python为非编译性语言，因此，Python源代码的主体并未被编译为可执行程序，仅将程序的启动项编译为了二进制可执行程序，大部分仍然采用源代码的方式发布。该系统的安装要求及安装方法详见下文。

**1.安装要求**

**1.1硬件环境**

处理器：主流的Intel或AMD处理器

内存容量：≥4GB

硬盘容量：≥2GB

**1.2软件环境**

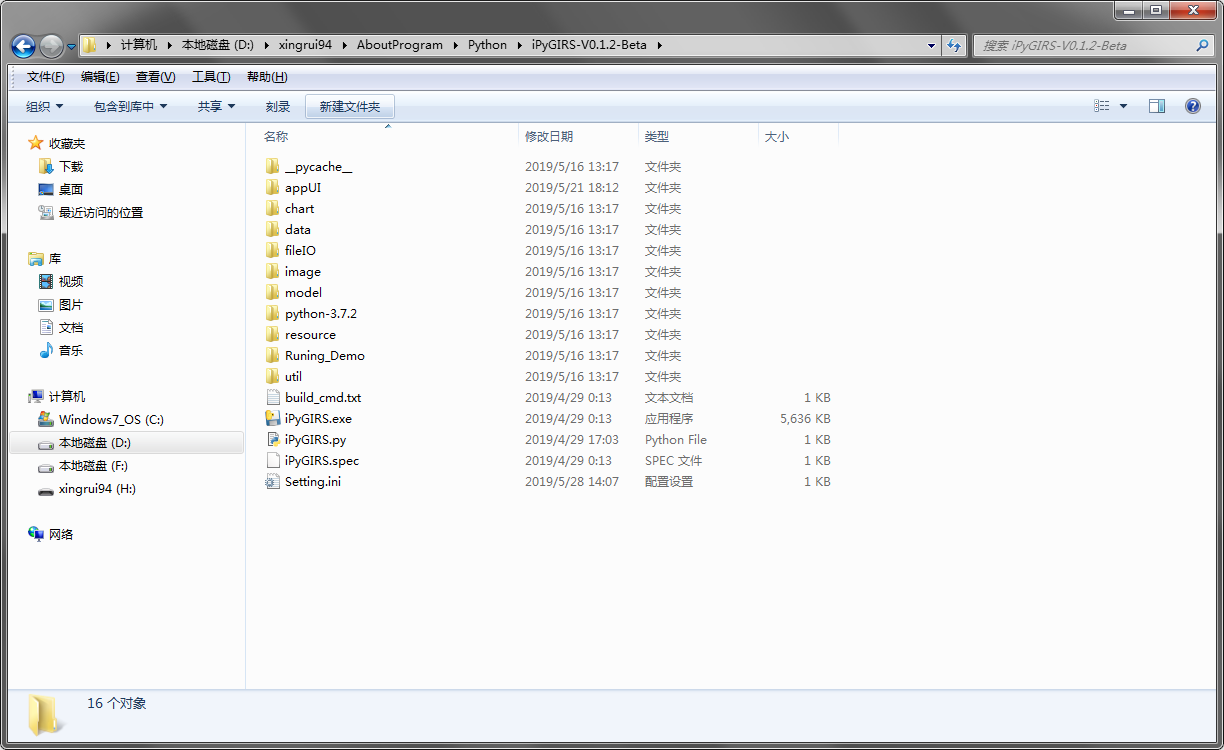
操作系统：Windows7 x64、Windows10 x64、Linux x64（未测试）

**2.安装方法**

如图2-1-1所示，解压“iPyGIRS-v0.1.2-Beta.zip”文件到自定义目录，即可得到如图2-1-2所示的程序文件，点击iPyGIRS.exe即可运行程序！



**图2-1-1 水环境遥感反演系统安装文件**



**图2-1-2 水环境遥感反演系统程序文件**

该系统安装过程的注意事项如下：

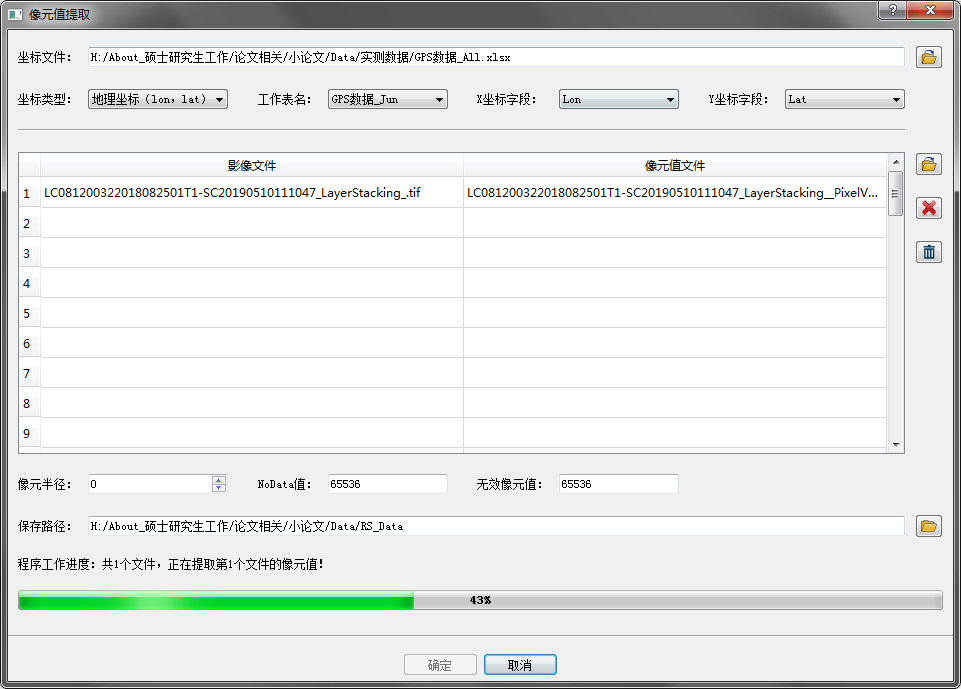
1. 解压的目标目录不能包含任何中文字符。
2. 为保证程序的正确运行，请勿更改解压后的文件（包括文件名、文件的相对路径）。

**第三章 系统使用说明**

**1.栅格数据工具**

**1.1像元值提取**

水质参数遥感反演是建立实际测量的水质数据和遥感影像光谱之间的关系模型的一个过程。因此提取像元值来构成建模样本是水质参数遥感反演的第一步。像元值提取，是指根据GPS坐标，提提遥感影像上对应位置的像元值。



**3-1-1像元值提取程序运行界面**

“像元值提取”程序的运行界面如图3-1-1所示，该程序的参数说明如表3-1-1所示：

**表3-1-1像元值提取程序参数使用说明**

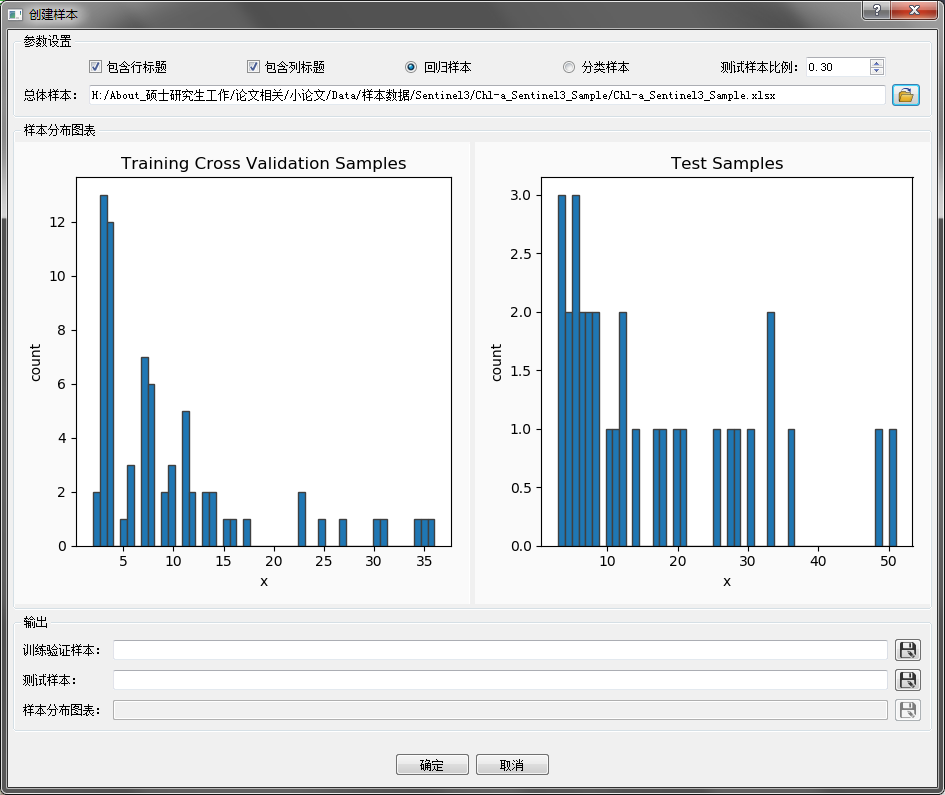
|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 坐标文件 | 该参数用以指定存储GPS坐标的Excel文件的绝对路径，可通过点击右侧的按钮进行选择。通过GPS定点得到GPS数据须要整理并保存为Excel文件，需要注意的是，Excel文件的格式只能是.xlsx格式，旧版本的.xls格式不受该程序的支持。此外，Excel文件中，数据应该包含列标题，否则程序可能无法正确识别经纬度坐标，或者将GPS坐标的第一条记录视为标题，从而导致少第一个点的像元值无法被提取。 |
| 坐标类型 | 该参数的默认值是地理坐标，也即由GPS仪器测量得到的经纬度坐标。该参数还支持投影坐标和图像坐标（也即像元的行列号）两种参数。 |
| 工作表名 | 该参数是指存储GPS坐标的Excel文件中的工作表的名称，一个Excel文件往往包含多个工作表（Sheet），因此，使用改程序时应当制定存储GPS坐标的工作表。 |
| X/Y坐标字段 | 该参数用以区分输入的水平方向（X）与垂直方向（Y）的坐标。Excel文件工作表中的坐标，包含在某两列中，X/Y是分别指X方向（如经度）和Y方向（如纬度）的坐标所在列的列标题。 |
| 像元半径 | 该参数是指提取像元值时的提取半径，当只提GPS坐标点的像元时，则设置该参数为0即可，当需要提取GPS坐标点及其最邻近的一个像元时，则设置该参数为1，依次类推，可根据实际需求进行设置。 |
| NoData值 | 该参数用以指定NoData像元的替代值。由于遥感影像会存在一些像元没有值或者是无效值的情况，因此设置该参数，可在提取像元值的过程中，将NoData像元的值替代为指定的值。 |
| 无效像元值 | 该参数用以指定无效像元的替代值。当输入的GPS坐标超出了影像的范围时，该GPS坐标对应的像元（当前图像不存在）即被视为无效像元。因此设置该参数，可在提取像元值的过程中，将无效GPS坐标对应的像元的值替代为指定的值。 |
| 保存路径 | 该参数用以指定输出文件的保存路径（磁盘目录）。“像元值提取”程序支持批量化的像元提取。即当坐标数据相同时，可输入多幅遥感影像，程序将自动命名每一幅影像对应输出的存储提取结果的Excel文件名。 |

**2.数据管理工具**

“数据管理工具”主要用以处理室外采集的表格类数据（如Excel文件、CSV文件、文本文件等），为水质参数反演模型的构建准备有效的样本数据集。该工具集包含“创建样本”、“归一化/标准化”、“特征处理”3个功能，其中，“特征处理”又包含“特征构建”和“特征选择”两个子功能。

**2.1创建样本**

无论是经验模型（统计模型），还是机器学习模型，都需要使用训练样本、验证样本和测试样本来分别完成训练模型、验证调试模型和模型泛化性评估工作。“创建样本”程序是一个将准备好的数据集划分为同分布的训练-验证集和测试集的工具。



**图3-2-1 创建样本程序运行界面**

“创建样本”程序的运行界面，如图3-2-1所示，当选择了总体样本文件后，程序的“样本分布图表”控件内将显示训练验证集和测试集的直方分布图，因此应该在打开总体样本文件之前，设置好相应参数。该程序的参数说明如表3-2-1所示：

**表3-2-1创建样本程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 包含行/列标题 | 该参数用以指定存储总体样本的Excel数据文件是否包含/列标题。由于行/列标题不作为样本数据的一部分，将不会被读取。错误的选项将导致多读或少读取样本数据。 |
| 行/列优先 | 该参数用以指定样本数据在Excel文件的工作表中是按行存储还是按列存储。按行存储即指每一行代表一个样本点，每一列（除最后一列之外）代表样本点的一个特征，列优先则与此相反。 |
| 测试样本比例 | 该参数用以指定测试样本占总体样本的比例，默认值是0.3，即30%，当指定测试样本比例后，训练-验证样本的比例为“1-测试样本比例”。 |
| 总体样本 | 该参数用以指定存储总体样本的Excel文件的绝对路径，可点击右侧按钮进行选择以打开总体样本文件。 |
| 训练验证样本 | 该参数用以指定保存训练-验证集的Excel文件的绝对路径，可点击右侧按钮进行选择以保存训练-验证样本集。 |
| 测试样本 | 该参数用以指定保存测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以保存测试样本集。 |
| 样本分布图表 | 该参数用以指定保存训练-验证集与测试集的分布图表的绝对路径，可点击右侧按钮进行选择以保存分布图（当前版本尚不支持此功能）。 |

**2.2归一化/标准化**

很多机器学习算法（如KNN、支持向量机、BP神经网络等）中的目标函数都是假设样本中的所有特征是零均值，并且同一阶数上的方差相同，因此，如果某个特征的方差比其他特征大几个数量级，那么它就会在学习算法中占据主导位置，导致学习器并不能像我们说期望的那样，从其他特征中学习。因此为了提高模型的精度和收敛速度，归一化或者标准化常常作为机器学习必要的预处理方式之一。该系统中，“归一化/标准化”程序的运行界面，如图3-2-2所示。



**图3-2-2 归一化/标准化程序运行界面**

归一化/标准化程序的参数说明如表3-2-2所示：

**表3-2-2归一化/标准化程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 包含行/列标题 | 该参数用以指定存储总体样本的Excel数据文件是否包含/列标题。由于行/列标题不作为样本数据的一部分，将不会被读取。错误的选项将导致多读或少读取样本数据。 |
| 行/列优先 | 该参数用以指定样本数据在Excel文件的工作表中是按行存储还是按列存储。按行存储即指每一行代表一个样本点，每一列（除最后一列之外）代表样本点的一个特征，列优先则与此相反。 |
| 数据文件 | 该参数用以指定存储训练-验证集或者测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以打开相应的数据文件。 |
| 归一化 | 选择该参数时，程序使用公式（1）所示，对数据进行归一化处理。 |
| 标准化 | 选择该参数时，程序使用公式（2）所示，对数据进行标准化处理。 |
| 最大/小值 | 该参数用以指定归一化处理的缩放范围，默认值为0-1。仅当数据预处理方式选择为归一化时，须设置该参数。 |
| 保存数据 | 该参数用以指定保存预处理完的训练-验证集或者测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以保存相应的数据文件。 |

…………………………………（1）

……………………………………………（2）

其中，μ为数据的均值，σ为方差。

**2.3特征处理**

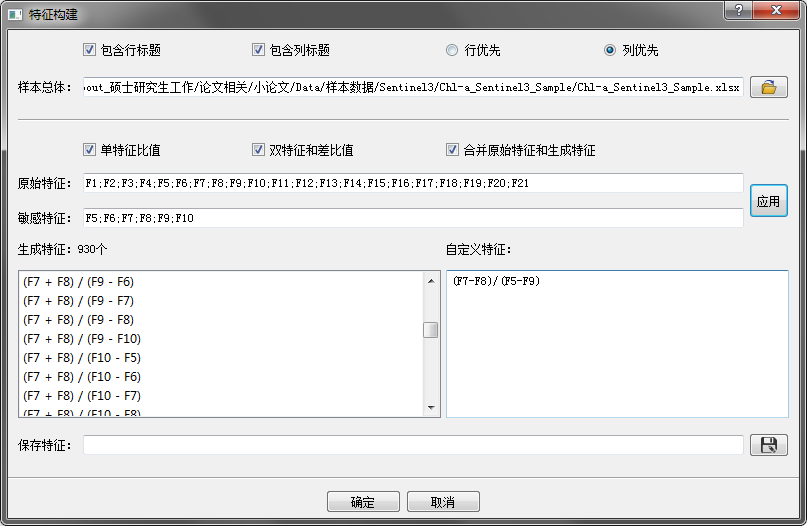
“数据决定了机器学习的上限，而算法只是尽可能逼近这个上限”，这里的数据指的就是经过特征工程得到的数据。特征工程是机器学习建模流程中必要的数据处理方式之一，特征工程指的是把原始数据转变为模型的训练数据的过程，它的目的就是获取更好的训练数据特征，使得机器学习模型逼近这个上限。特征工程能使得模型的性能得到提升，有时甚至在简单的模型上也能取得不错的效果。特征工程在机器学习中占有非常重要的作用，一般认为括特征构建、特征提取、特征选择三个部分。

**2.3.1特征构建**

“特征构建”程序是一个在准备好的数据集的已有特征基础上，根据指定的计算方式，计算以构建新特征的工具。该程序的各个参数作用及设置说明如表3-2-3所示，程序运行界面如图3-2-3所示：

**表3-2-3特征构建程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 包含行/列标题 | 该参数用以指定存储总体样本的Excel数据文件是否包含/列标题。由于行/列标题不作为样本数据的一部分，将不会被读取。错误的选项将导致多读或少读取样本数据。 |
| 行/列优先 | 该参数用以指定样本数据在Excel文件的工作表中是按行存储还是按列存储。按行存储即指每一行代表一个样本点，每一列（除最后一列之外）代表样本点的一个特征，列优先则与此相反。 |
| 样本总体 | 该参数用以指定存储训练-验证集或者测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以打开相应的数据文件。 |
| 单特征比值 | 勾选该参数，程序会自动计算任意两个敏感特征之间的比值，如：F1/F2 |
| 双特征和差比值 | 勾选该参数，程序会自动计算任意两个敏感特征之间的和/差与差/和的比值，即归一化比值计算，如：（F1+F2）/（F1-F2） |
| 合并原始特征和生成特征 | 勾选该参数，程序保存构建特征时，会将原始特征与生成特征保存到同一个Excel文件。 |
| 自定义特征 | 程序会自动根据文本框内输入的公式进行特征构建。 |



**图3-2-3 特征构建程序运行界面**

**2.3.2特征选择**

对于机器学习模型，在有限的样本量情况下，特征过多容易导致模型过拟合，“特征选择”程序根据样本数据的各特征方差、特征与标签之间的皮尔逊相关系数和距离相关系数，用以提取对模型有用的特征信息，提高模型的泛化性能。该程序的各参数作用及设置说明如表3-2-4所示，程序运行界面如图3-2-4

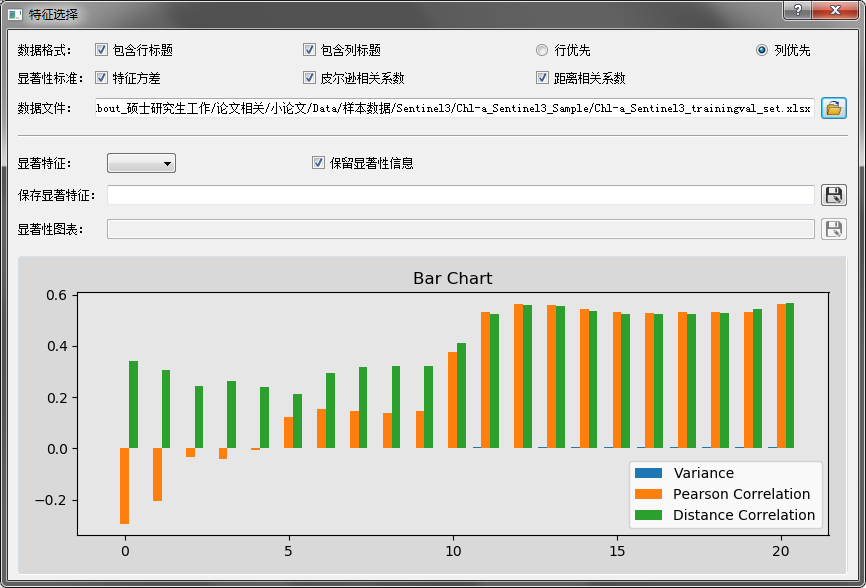
**表3-2-4特征选择程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 包含行/列标题 | 该参数用以指定存储总体样本的Excel数据文件是否包含/列标题。由于行/列标题不作为样本数据的一部分，将不会被读取。错误的选项将导致多读或少读取样本数据。 |
| 行/列优先 | 该参数用以指定样本数据在Excel文件的工作表中是按行存储还是按列存储。按行存储即指每一行代表一个样本点，每一列（除最后一列之外）代表样本点的一个特征，列优先则与此相反。 |
| 数据文件 | 该参数用以指定存储训练-验证集或者测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以打开相应的数据文件。 |
| 特征方差 | 选择该参数时，程序使用公式（1），计算样本数据各个特征的方差。 |
| 皮尔逊相关系数 | 选择该参数时，程序使用公式（2），计算样本数据各个特征与标签数据之间的皮尔逊相关系数。比值，即归一化比值计算，如：（F1+F2）/（F1-F2） |
| 距离相关系数 | 选择该参数时，程序使用公式（3），计算样本数据各个特征与标签数据之间的距离相关系数。 |
| 显著特征 | 用户可根据“特征方差”、“皮尔逊相关系数”和“距离相关系数”三个标准，在该参数中勾选出较为显著的特征。 |
| 保留显著特征信息 | 勾选该参数，程序将各个特征的“特征方差”、“皮尔逊相关系数”和“距离相关系数”的计算结果同时保存到Excel中去。 |
| 保存显著特征 | 该参数用以指定保存显著特征数据的Excel文件的绝对路径，可点击右侧按钮进行选择以保存只包含显著特征的样本集。 |
| 显著特征图表 | 该参数用以指定保存特征显著性信息图表的绝对路径，可点击右侧按钮进行选择以保存图表（当前版本尚不支持此功能） |

……………………………………（1）

…………………………（2）

…………………（3）



**图3-2-3 特征选择程序运行界面**

**3.机器学习工具**

“机器学习工具”是《水环境遥感反演》系统的核心模块，主要实现了遥感反演的机器学习模型建模功能，实现了“线性回归”、“支持向量机”、“BP神经网络”、“随机森林”、“LightGBM”五个回归模型。此外，该工具箱内还包含了“应用模型”功能，用以实现遥感影像的反演。

**3.1回归模型**

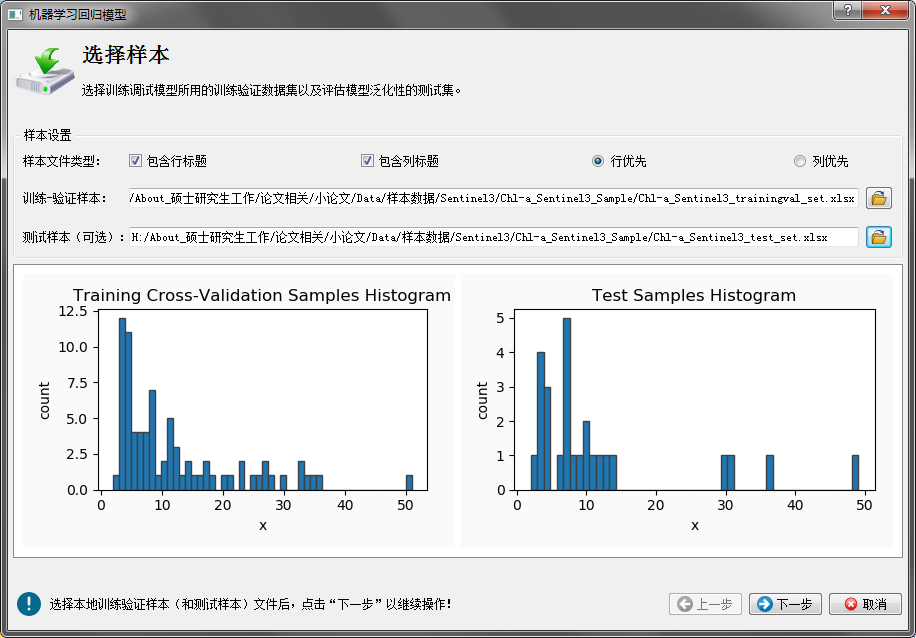
回归模型（regression model）程序的操作逻辑与Matlab中神经网络工具箱类似，一共包括5个操作步骤，每个步骤对应着不同的程序界面。各个界面的操作参数见下文。

**3.1.1选择样本**

“选择样本”程序运行界面如表3-3-1所示，具体的各项参数作用及设置说明如下：

**表3-3-1选择样本程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 包含行/列标题 | 该参数用以指定存储总体样本的Excel数据文件是否包含/列标题。由于行/列标题不作为样本数据的一部分，将不会被读取。错误的选项将导致多读或少读取样本数据。 |
| 行/列优先 | 该参数用以指定样本数据在Excel文件的工作表中是按行存储还是按列存储。按行存储即指每一行代表一个样本点，每一列（除最后一列之外）代表样本点的一个特征，列优先则与此相反。 |
| 训练-验证样本 | 该参数用以指定保存训练-验证集的Excel文件的绝对路径，可点击右侧按钮进行选择以打开训练-验证样本集。 |
| 测试样本（可选） | 该参数用以指定保存测试集的Excel文件的绝对路径，可点击右侧按钮进行选择以打开测试样本集。该参数为可选参数，如果保持默认的空值，点击“下一步”操作时，程序会发出如的询问提示！ |



**图3-3-1 选择样本程序运行界面**

**3.1.2设置参数**

“设置参数”程序界面如图3-3-2所示，该界面的功能是，选择模型并设置模型的基本参数和调试参数。机器学习工具的“回归模型”实现了“线性回归”、“支持向量机”、“BP神经网络”、“随机森林”、“LightGBM”5种回归模型。



**图3-3-2 设置参数程序运行界面**

**3.1.3训练模型**

“训练模型”程序界面如图3-3-3所示，该界面的功能是根据“选择样本”和“设置参数”两个界面的数据，训练并测试（当且仅当“选择样本”界面输入了测试样本）模型。其中需要设置“交叉验证”和“工作核心数”两个模型训练参数，参数的作用和及设置说明如表3-3-2所示：

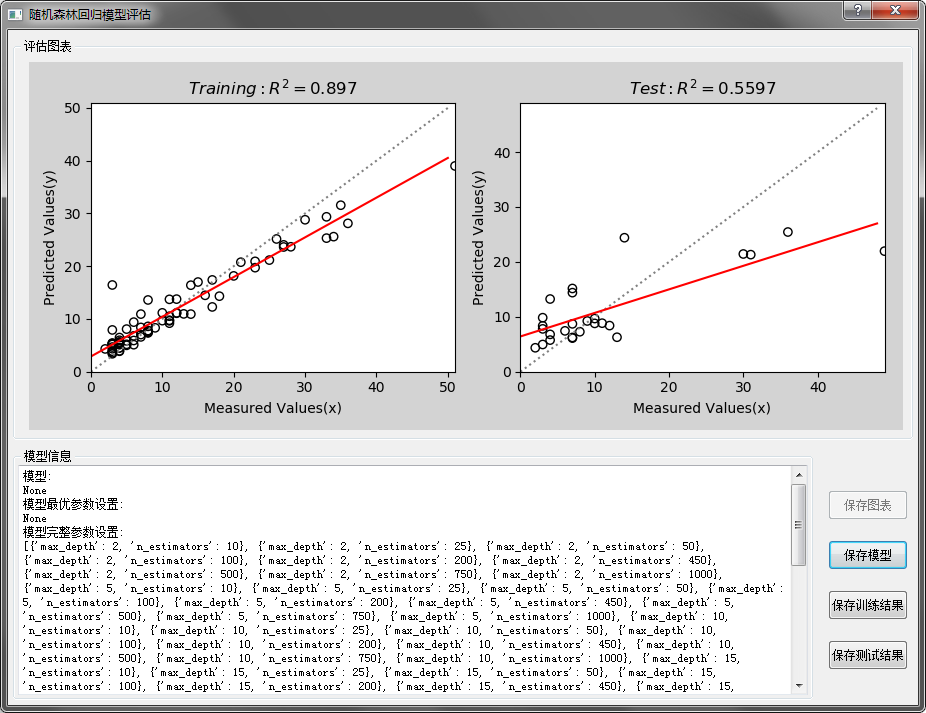
**表3-3-2训练模型程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 交叉验证 | 该参数用以指定机器学习模型训练-验证过程中K折交叉验证的K值。默认值为10，即训练集将被划分为10分，训练过程中每次使用9份子数据作为训练集进行模型训练，剩下的1份用以模型的验证，评估当前参数设置下的模型得分，以此类推，共进行10次训练-验证过程。 |
| 工作核心数 | 该参数用以指定机器学习模型训练-验证，测试、应用过程中使用的CPU核心数，使用核心数越多，模型的计算速度将越快，该参数默认值为2。 |



**图3-3-3 训练模型程序运行界面**

“训练模型”程序界面上的“结果”控件在模型训练-验证结束后，会显示相应的模型得分，此时，可点击“绘制模型评估”图表以查看回归模型的得分散点图，如图3-3-4所示。点击“绘制训练信息图表”将显示模型的学习曲线图（该功能在当前版本中尚未实现）。



**图3-3-4 训练模型程序运行界面**

**3.1.4优化模型**

“优化模型”程序界面如图3-3-5所示，具体的参数作用及设置说明，见程序界面说明。



**图3-3-5 优化模型程序运行界面**

**3.1.5导出数据**

“导出数据”程序界面如图3-3-6所示，可勾选所要保存的数据，并在文本框内，自定义名称。参数“保存路径”是必填参数，可点击右侧的按钮以选择保存路径，选择完保存路径后，仍需点击“保存结果”按钮，才能将数据保存到磁盘上的指定位置。“重置参数”按钮将所有导出数据选项及名称恢复为程序的默认值。



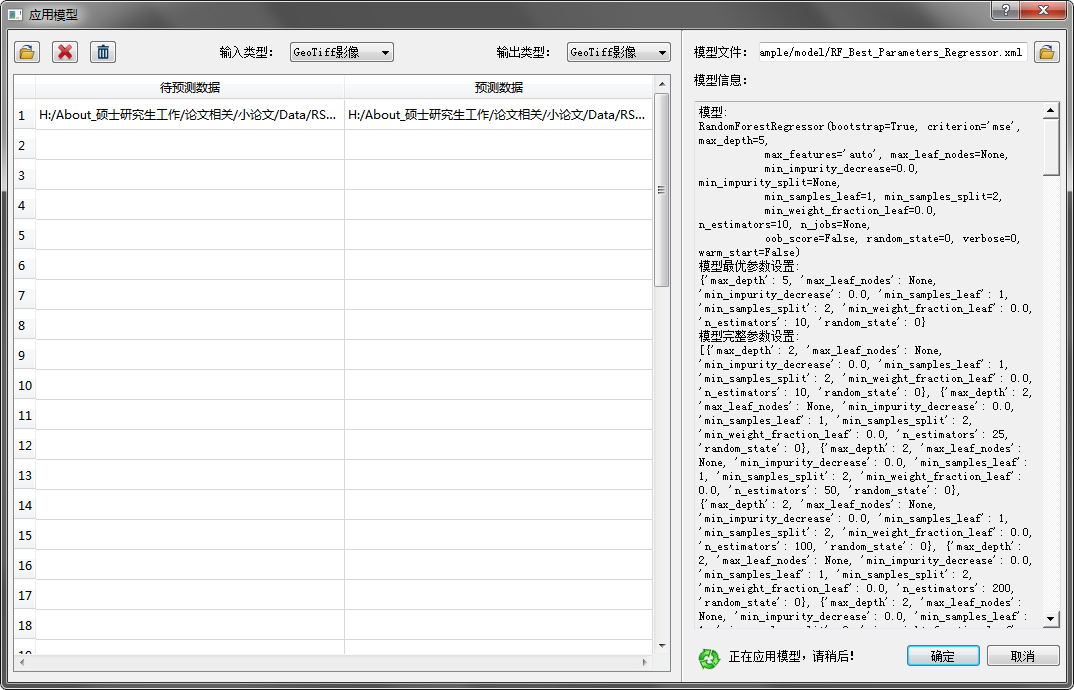
**图3-3-6 导出数据程序运行界面**

**3.2应用模型**

“应用模型”程序，使用训练好的模型将符合模型要求的待预测数据计算出来。该程序的参数租用及设置说明如下：

**表3-3-3应用模型程序参数使用说明**

|  |  |
| --- | --- |
| **参数** | **使用说明** |
| 输入/输出类型 | 在《水环境遥感反演系统》中，该参数用以指定程序的待预测数据和预测数据只能为Geotiff影像。 |
| 模型文件 | 该参数用以指定保存训练好的机器学习模型文件的绝对路径，可点击右侧按钮进行选择以打开测试样本集。须要注意的是，模型文件包括.xml和.pkl两个，.xml作为该程序的输入，主要用以保存模型的名称、特征信息、pkl文件信息等。 |



**图3-3-7 导出数据程序运行界面**