数据挖掘-随机森林

概述

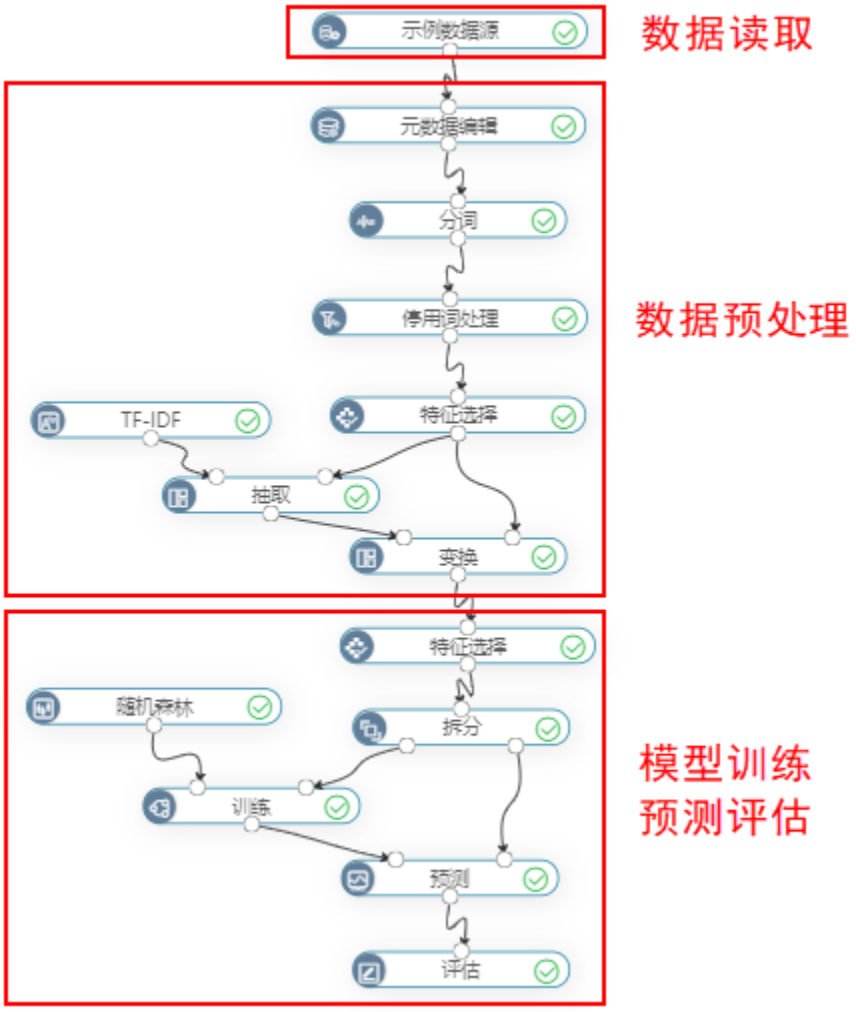
随机森林指的是利用多棵树构成森林对样本进行训练并预测的一种分类器。但是每棵决策树之间没有关联，每棵树都是基于随机抽取的样本和特征进行独立训练。

随机森林算法广泛应用于分类问题。其是决策树的组合，将许多决策树联合到一起，以降低过拟合的风险。随机森林支持连续数据或离散数据进行二分类或多分类。

优势：可反映出特征重要性。

示例

使用“垃圾短信识别”案例数据，预测是否为垃圾短信。



其中，分词是为了将短信文本进行分成词语方便分析；停用词处理是为了去除不必要的词语、标点符号、语气词等；TF-IDF是为了计算文本数据的idf值，方便进入模型训练。

随机森林的参数如下：

参数名称	值	说明
分裂特征的数量	取值范围：>=2的整数； 默认值：32。	对连续类型特征进行离散时的分箱数； 该值越大，模型会计算更多连续型特征分裂点且会找到更好的分裂点，但同时也会增加模型的计算量；
树的深度	取值范围：[1,30]的整数；默认值为4。	当模型达到该深度时停止分裂； 树的深度越大，模型训练的准确度更高，但同时也会增加模型的计算量且会导致过拟合；

树的个数	取值范围：>=1的整数。默认值为20。	随机森林中决策树的棵数。
衡量准则	gini	裂分标准，Entropy表示熵值，Gini表示基尼指数；
	entropy	