

数据挖掘-随机森林特征选择

概述

特征选择是为算法服务的，选择不同的特征会直接影响到模型的效果。随机森林特征行选择，就是使用随机森林算法，来自动选择相关性高的特征。

它是一种基于随机森林的封装式特征选择算RFFS，以随机森林算法为基本工具，以分类或回归精度作为准则函数,采用序列后向选择和广义序列后向选择方法进行特征选择。

- 概述
- 输入/输出
- 参数设置
- 示例

输入/输出

输入	一个输入端口，用于接收前置节点传下来的数据集。
输出	一个输出端口，用于输出增加了离散后的字段的数据集。

参数设置

参数名称	说明	备注
选择特征列	选择需要的特征列，必须是数值列	必填（特征列中不能含有null）
选择标签列	选择做为标签列的字段	必填
需选择的特征数量	从待选择的特征列中输出特征列的数量	默认值为1, 范围是[1, 已选择特征的数量]的整数
选择方法	分类	
	回归	
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	

示例

使用“鸢尾花数据”，预测鸢尾花类别。选取4个特征列和1个标签列，需选择的特征数量为3，选择方法分类，衡量准则为gnin，其他参数默认。结果输出“featureSelector”列，表示为从4个特征列的值进行特征选择相关性较高的3个特征。如下图：

al_Width	Petal_Length	Petal_Width	Species	SpeciesIndex	featureSelector
3.5	1.4	0.2	setosa	2.0	[5.1,3.5,1.4,0.2]
3.0	1.4	0.2	setosa	2.0	[4.9,3.0,1.4,0.2]
3.2	1.3	0.2	setosa	2.0	[4.7,3.2,1.3,0.2]
3.1	1.5	0.2	setosa	2.0	[4.6,3.1,1.5,0.2]
3.6	1.4	0.2	setosa	2.0	[5.0,3.6,1.4,0.2]
3.9	1.7	0.4	setosa	2.0	[5.4,3.9,1.7,0.4]
3.4	1.4	0.3	setosa	2.0	[4.6,3.4,1.4,0.3]
3.4	1.5	0.2	setosa	2.0	[5.0,3.4,1.5,0.2]
2.9	1.4	0.2	setosa	2.0	[4.4,2.9,1.4,0.2]
3.1	1.5	0.1	setosa	2.0	[4.9,3.1,1.5,0.1]
3.7	1.5	0.2	setosa	2.0	[5.4,3.7,1.5,0.2]

点击鼠标右键查看分析结果：

从4个特征列中选取3个相关性最高的特征进行展示。如下图：

