

Smartbi V10.5.8-数据挖掘

注意：（新特性列表中：+表示新增；^表示增强）

- 具体改进点如下：
- +【数据挖掘】支持异常检测使用场景功能(CBLOF)
- +【数据挖掘】支持异常检测使用场景功能(孤立森林)
- +【数据挖掘】增强特征工程(WOE)模型保存
- +【数据挖掘】支持自定义调整节点目录结构
- ^【数据挖掘】关联规则和聚类支持模型自学习
- ^【数据挖掘】评估报告重构
- ^【ETL】指标模型维度和事实表新建自助ETL时，自动添加目标数据源节点

具体改进点如下：

新增	增强
+【数据挖掘】支持异常检测使用场景功能(CBLOF)	^【数据挖掘】关联规则和聚类支持模型自学习
+【数据挖掘】支持异常检测使用场景功能(孤立森林)	^【数据挖掘】评估报告重构
+【数据挖掘】增强特征工程(WOE)模型保存	^【ETL】指标模型维度和事实表新建自助ETL时，自动添加目标数据源节点
+【数据挖掘】支持自定义调整节点目录结构	

+【数据挖掘】支持异常检测使用场景功能(CBLOF)

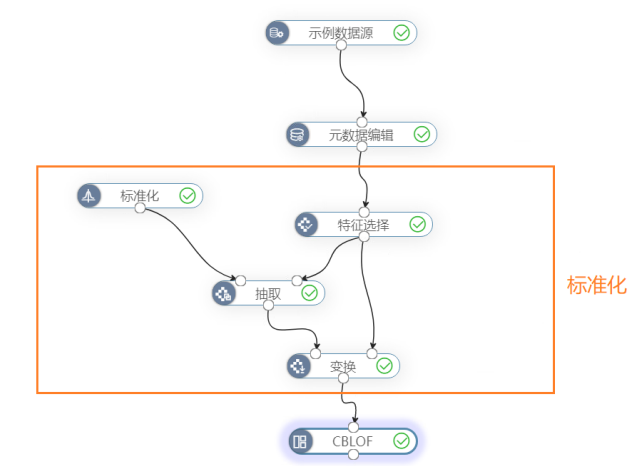
背景介绍

异常检测广泛应用在需要发现离群点的场景中如：欺诈检测、入侵检测等，在数据挖掘中属于经典的场景之一，因此将部分较为经典的异常检测算法通过spark进行分布式实现。

功能简介

CBLOF（Cluster-Based Local Outlier Factor，基于聚类的本地异常因子）是一个异常检测节点，原理是先用聚类算法把数据分为K个簇，而后通过设定占比阈值和突降倍数阈值，把簇区分为大簇和小簇，聚类完成后，计算每个点到最邻近大簇的距离（邻近距离），邻近距离越大的数据点为异常数据的概率越大。

CBLOF算法适用于当没有已知正常的的数据时，对所有输入的新数据进行异常值的辨别。



参考文档

详情请参考 [数据挖掘 - CBLOF](#)

+【数据挖掘】支持异常检测使用场景功能(孤立森林)

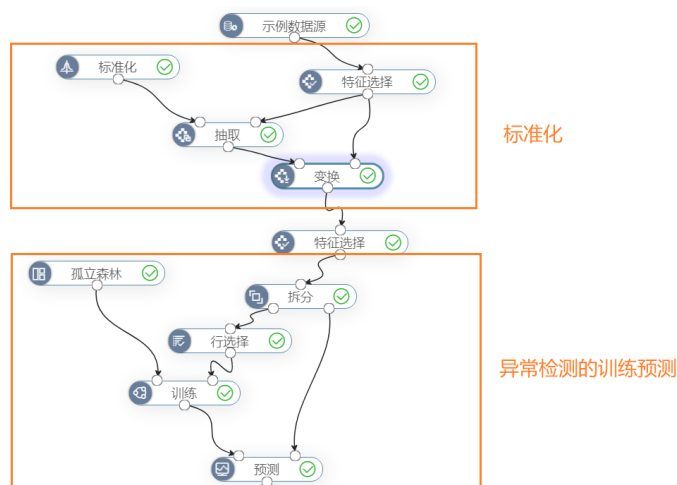
背景介绍

异常检测广泛应用在需要发现离群点的场景中如：欺诈检测、入侵检测等，在数据挖掘中属于经典的场景之一，因此将部分较为经典的异常检测算法通过spark进行分布式实现。

功能简介

孤立森林是一个异常检测算法节点，需要配合特征选择、训练、预测节点使用（异常检测算法为无监督学习，不需要评估节点）。孤立森林对正常数据进行采样，训练时随机生成决策树对数据进行划分。在预测未知数据时，决策树划分出某单个数据点需要的划分次数越少，对应的异常分数越高，该数据点为异常数据的概率越大。

孤立森林适用于用已知为正常的的数据作为训练数据，然后对未知的新数据作预测，检测新数据中的异常数据。



参考文档

详情请参考 [数据挖掘 - 孤立森林](#)。

+【数据挖掘】增强特征工程(WOE)模型保存

背景介绍

旧的特征工程功能没有保存模型机制，导致用户需要手工记住参数来进行模型发布，因此新增支持类似WOE特征工程模型的保存机制。

功能简介

WOE编码节点和评分卡构建节点现在支持保存模型，保存的模型可以展示模型参数，以及再次使用进行预测。

新增了数据预测节点用于对特征工程模型（WOE编码模型、评分卡模型）进行预测，可以配合这些模型的保存功能，把已保存的模型接入数据预测节点对新数据进行预测。

参考文档

详情请参考 [数据挖掘 - 数据预测](#)。

+【数据挖掘】支持自定义调整节点目录结构

背景介绍

用户有时候想要调整挖掘实验里左侧节点树的目录结构。同时挖掘能够上传自定义节点，但是无法修改目录结构，导致上传的新节点有时候无法正确归类，因此需要能够支持自定义目录结构功能。

功能简介

用户可以根据需要以拖拽的方式调整节点目录中的文件夹顺序、节点顺序、节点归类等。上传的自定义节点也能在目录结构中调整。支持新建目录。



参考文档

详情请参考 [数据挖掘 - 自定义节点目录结构](#)。

^【数据挖掘】关联规则和聚类支持模型自学习

背景介绍

旧的模型自学习功能只支持分类算法的自学习任务，无法作用于聚类算法和关联规则算法，因此模型自学习需要支持关联规则和聚类任务。

功能简介

可以在聚类任务和关联规则任务实验中进行模型自学习的设置，关联到对应的聚类模型和关联规则模型，实现对该模型的自动更新。

①自动训练设置

选择评估节点 *

聚类评估

监控指标 *

SSE

上线阈值 *

2000

注意回归中mas,rmse,mse和聚类中SSE需小于阈值才会上线

预警阈值 *

0.3

选择训练节点 *

聚类训练

更新服务/预测 *

葡萄预测

已关联自学习任务的服务和预测不可选

模型节点 *

葡萄预测

上线方式

☒ 自动上线

☐ 手动上线

取消

保存并设置定时任务

参考文档

详情请参考 [数据挖掘 - 模型自学习和模型批量预测使用示例](#)。

^【数据挖掘】评估报告重构

背景介绍

旧的评估报告的功能通过电子表格的方式展现，重构后通过仪表盘的方式展现。

功能简介

通过在数据模型中建有若干个Java查询，用于获取挖掘实验中的评估节点中的数据，获取的结果输出到仪表盘，并最终能够通过仪表盘进行数据展现。



参考文档

详情请参考 [数据挖掘 - 模型对比](#)。

【ETL】指标模型维度和事实表新建自助ETL时，自动添加目标数据源节点

背景介绍

指标模型基于维度和事实表新建自助ETL时，自动添加目标数据源节点，并填写好相关参数。

功能简介

指标模型基于维度和事实表新建自助ETL时，会在新建的ETL中自动添加关系目标表（覆盖）节点，并自动填写数据库、Schema、表名的信息。

