Smartbi V10-数据挖掘

注意：（新特性列表中：+表示新增；^表示增强）

- +【数据挖掘】新增自助机器学习，能够快速创建挖掘实验
- +【数据挖掘】新增Kafka数据源节点
- +【自助ETL/数据挖掘】数据源新增Excel文件
- +【自助ETL/数据挖掘】目标源支持GreenPlum数据库
- +【数据挖掘】数据预处理增加下采样节点
- +【数据挖掘】新增SMOTE数据预处理方式
- +【自助ETL/数据挖掘】数据预处理新增值替换节点
- +【数据挖掘】特征工程新增GBDT特征选择节点
- +【数据挖掘】统计分析支持高维数据可视化
- +【数据挖掘】评分卡分析新增PSI评估节点
- +【数据挖掘】文本分析增加词向量节点
- +【数据挖掘】新增聚类评估节点，用于呈现聚类算法常见评价指标值
- ^【自助ETL/数据挖掘】关系数据源支持参数设置
- ^【自助ETL/数据挖掘】关系数据源支持分区设置，提升数据抽取效率
- ^【数据挖掘】关系目标表（追加）节点追加数据前支持删除表中数据
- ^【自助ETL/数据挖掘】元数据编辑支持修改原字段名及顺序
- ^【数据挖掘】派生列、聚合、全表统计节点新增多个函数
- ^【数据挖掘】分词节点新增自定义全局词典和分词算法
- ^【数据挖掘】完善Python算法节点功能
- ^【数据挖掘】查看输出支持预览数据导出到本地
- ^【自助ETL/数据挖掘】查看输出增加列筛选项
- ^【数据挖掘】节点输出字段支持排序
- ^【数据挖掘】增强整个页面的操作
- ^【自助ETL/数据挖掘】支持缓存节点数据，减少执行实验等待时间
- ^【自助ETL/数据挖掘】支持多节点分组收缩和展开
- <【自助ETL/数据挖掘】关系目标源拆分为追加、覆盖、插入或更新数据节点
- <【数据挖掘】拆分归一化算法为多个节点

具体改进点如下：

新增	增强	变更

+【数据挖掘】新增自助机器学习，能够快速创建挖掘实验 +【数据挖掘】新增Kafka数据源节点 +【自助ETL/数据挖掘】数据源新增Excel文件 +【自助ETL/数据挖掘】目标源支持GreenPlum数据库 +【数据挖掘】数据预处理增加下采样节点 +【数据挖掘】新增SMOTE数据预处理方式 +【自助ETL/数据挖掘】数据预处理新增值替换节点 +【数据挖掘】特征工程新增GBDT特征选择节点 +【数据挖掘】统计分析支持高维数据可视化 +【数据挖掘】评分卡分析新增PSI评估节点 +【数据挖掘】文本分析增加词向量节点 +【数据挖掘】新增聚类评估节点，用于呈现聚类算法常见评价指标值	^【自助ETL/数据挖掘】关系数据源支持参数设置 ^【自助ETL/数据挖掘】关系数据源支持分区设置，提升数据抽取效率 ^【数据挖掘】关系目标表（追加）节点追加数据前支持删除表中数据 ^【自助ETL/数据挖掘】元数据编辑支持修改原字段名及顺序 ^【数据挖掘】派生列、聚合、全表统计节点新增多个函数 ^【数据挖掘】分词节点新增自定义全局词典和分词算法 ^【数据挖掘】完善Python算法节点功能 ^【数据挖掘】查看输出支持预览数据导出到本地 ^【自助ETL/数据挖掘】查看输出增加列筛选项 ^【数据挖掘】节点输出字段支持排序 ^【数据挖掘】增强整个页面的操作 ^【自助ETL/数据挖掘】支持缓存节点数据，减少执行实验等待时间 ^【自助ETL/数据挖掘】支持多节点分组收缩和展开	<【自助ETL/数据挖掘】关系目标源拆分为追加、覆盖、插入或更新数据节点 <【数据挖掘】拆分归一化算法为多个节点

+【数据挖掘】新增自助机器学习，能够快速创建挖掘实验

背景介绍

随着数据挖掘在各种领域不断被应用，越来越多的人开始使用机器学习，而使用机器学习不仅需要用户具备一定专业知识，还需要花费大量的精力来进行算法与模型的选择。为了进一步降低用户的使用门槛，我们在数据挖掘中，支持使用自助机器学习功能快速创建数据挖掘实验，能够自动化的完成更多的工作，也能让没有太多专业知识的人也能使用机器学习。

功能简介

新建回归、分类或聚类实验时，只需配置数据源、算法、特征的设置项，系统可快速自动生成实验。

参考文档

关于AutoML的功能，详情请参考 数据挖掘-自助机器学习 。

+【数据挖掘】新增Kafka数据源节点

背景介绍

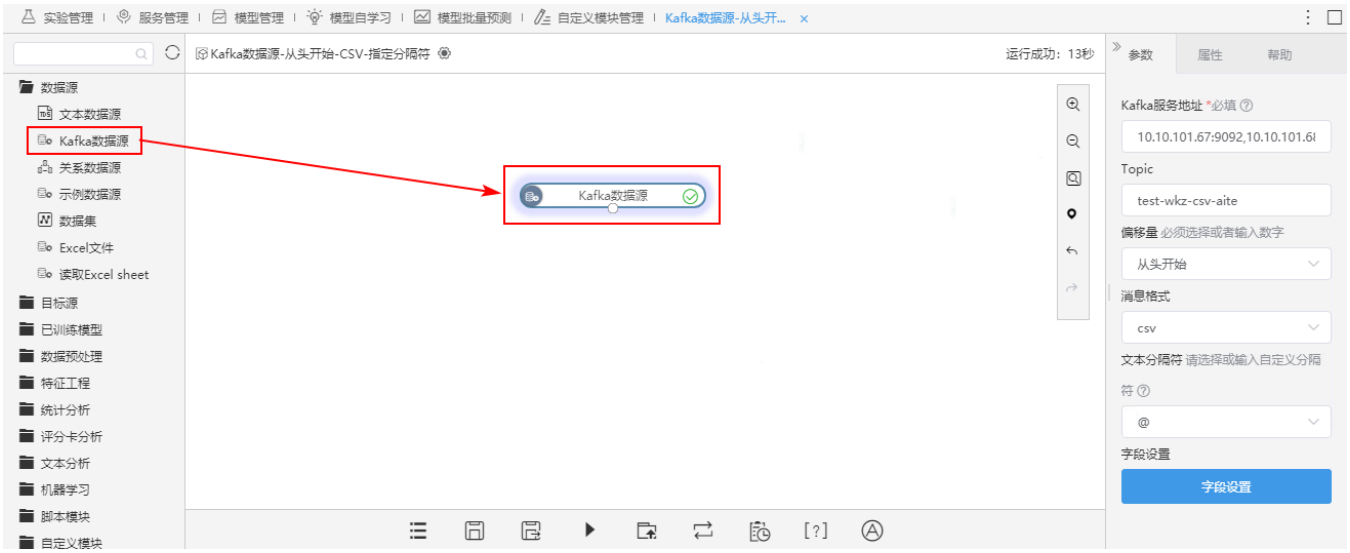
Kafka是一种高吞吐量的分布式发布订阅消息系统，经常用于实时流数据架构，提供实时分析。它具有高吞吐量、低延迟，每秒可以处理几万条消息，延迟最低只有几毫秒，以及可扩展性、持久性、可靠性、容错性、高并发等优点。因此，Smartbi在V10版本新增了Kafka数据源。

功能简介

Kafka作为数据源一般用来缓存数据，然后由Storm消费Kafka中的数据进行实时处理，有以下三种使用场景：

- 准实时的数据处理：通过任务调度，持续消费Kafka中的数据，提供给一系列数据处理节点进行处理，处理后的结果可以输出到目标数据库；
- 模型自学习：通过任务调度，持续消费Kafka中的数据进行模型自学习；
- 模型批量预测：通过任务调度，定时消费Kafka中的数据进行批量预测。

新增的Kafka数据源如图：



详情参考

关于Kafka数据源，详情请参考 数据挖掘-数据的输入和输出 。

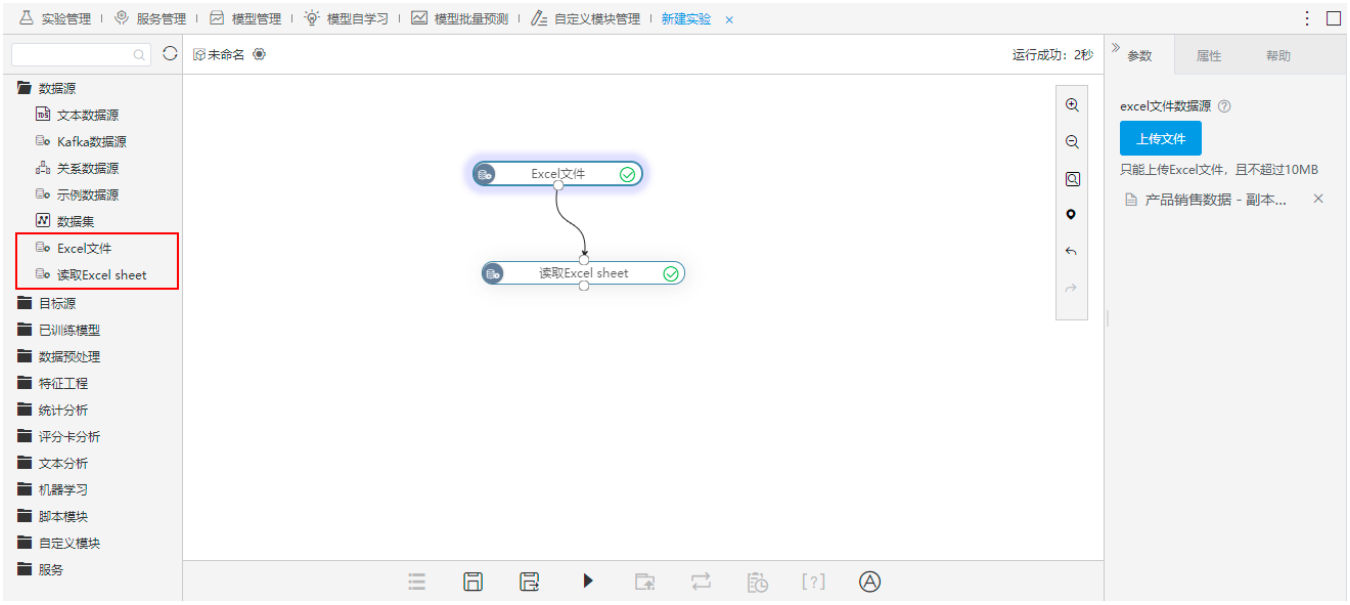
+【自助ETL/数据挖掘】数据源新增Excel文件

背景介绍

在实际应用中，不同的用户有着不同的数据导入需求，有的用户想要通过导入Excel数据文件的方式修改表结构等。为了满足用户需求，V10版本我们在自助ETL和数据挖掘中，新增Excel文件数据源，可通过上传Excel文件的方式导入数据，丰富了数据来源。

功能简介

V10版本，在自助ETL和数据挖掘的数据源节点中，新增Excel文件、读取Excel文件sheet文件节点，可通过上传Excel文件的方式导入需要的数据。



详情参考

关于Excel文件节点、读取Excel文件sheet文件节点功能，详情请参考 Excel文件数据源 。

+【自助ETL/数据挖掘】目标源支持GreenPlum数据库

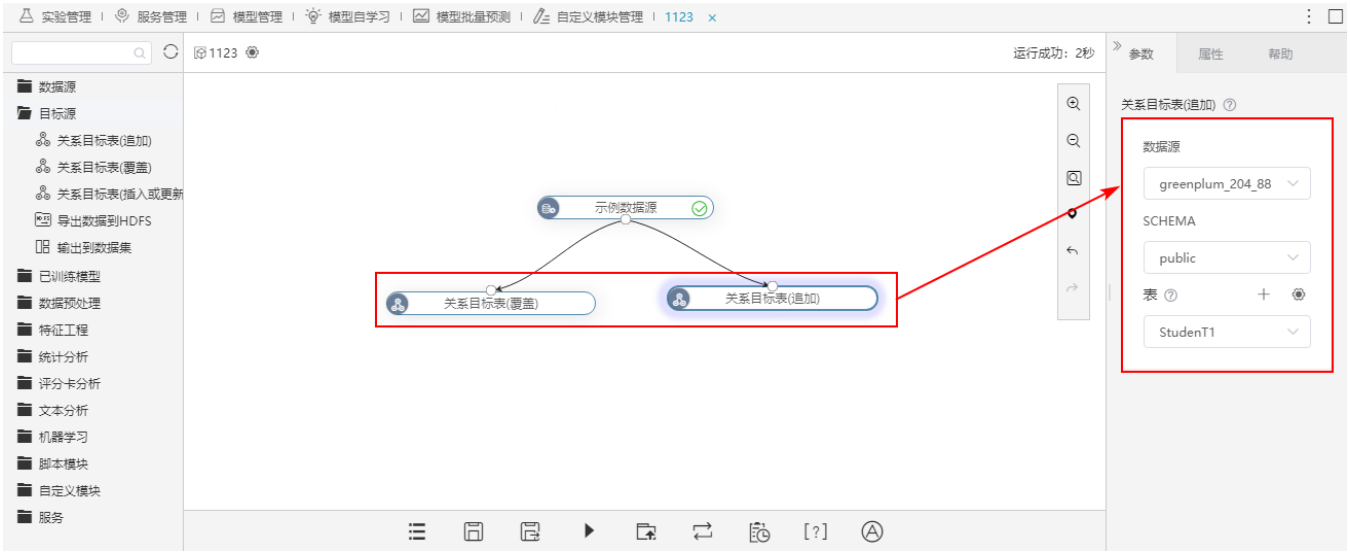
背景介绍

随着数据的爆炸性增长，用户对存储数据量的需求不断增加，产品在数据挖掘和自助ETL中，关系目标表（追加）和关系目标表（覆盖）节点支持使用GreenPlum数据库。

GreenPlum是一个面向数据仓库应用的关系型数据库，因为有良好的体系结构，所以在数据存储、高并发、高可用、线性扩展、反应速度、易用性和性价比等方面都有非常明显的优势，同时配置简单，因此深受用户的欢迎。

功能简介

在数据挖掘和自助ETL中，关系目标表（追加）和关系目标表（覆盖）节点支持GreenPlum数据库。



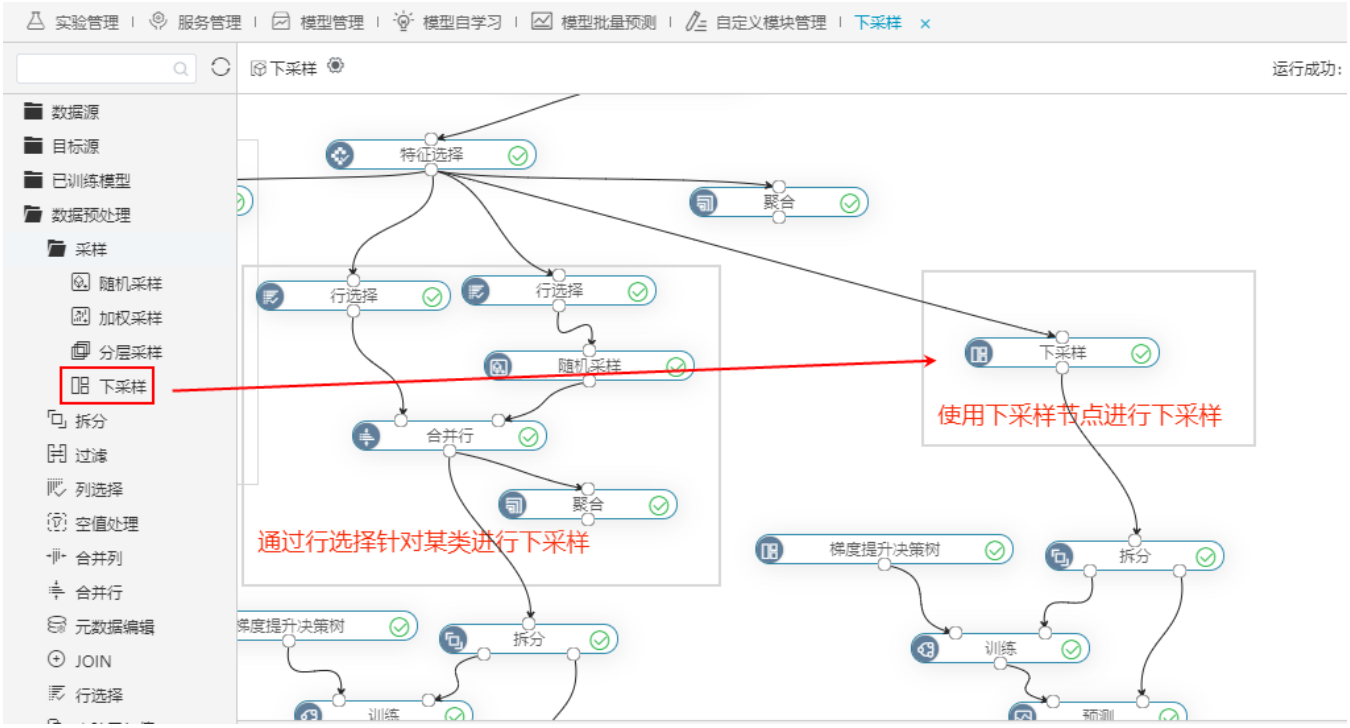
+【数据挖掘】数据预处理增加下采样节点

背景介绍

在数据挖掘过程中，原始数据的不均匀分布会影响到数据特征抽取或模型学习数据特征的效果，出现错判的情况。V10版本新增下采样节点，可对原始数据进行初步加工，对出现频次较高的数据按照一定规则抽取一定数据使得整体分布均匀。

功能简介

新增下采样节点，可通过移除数据量较多类别的部分数据，使样本达到均衡。



参考文档

关于数据挖掘的下采样节点，详情请参考　　数据挖掘-采样　。

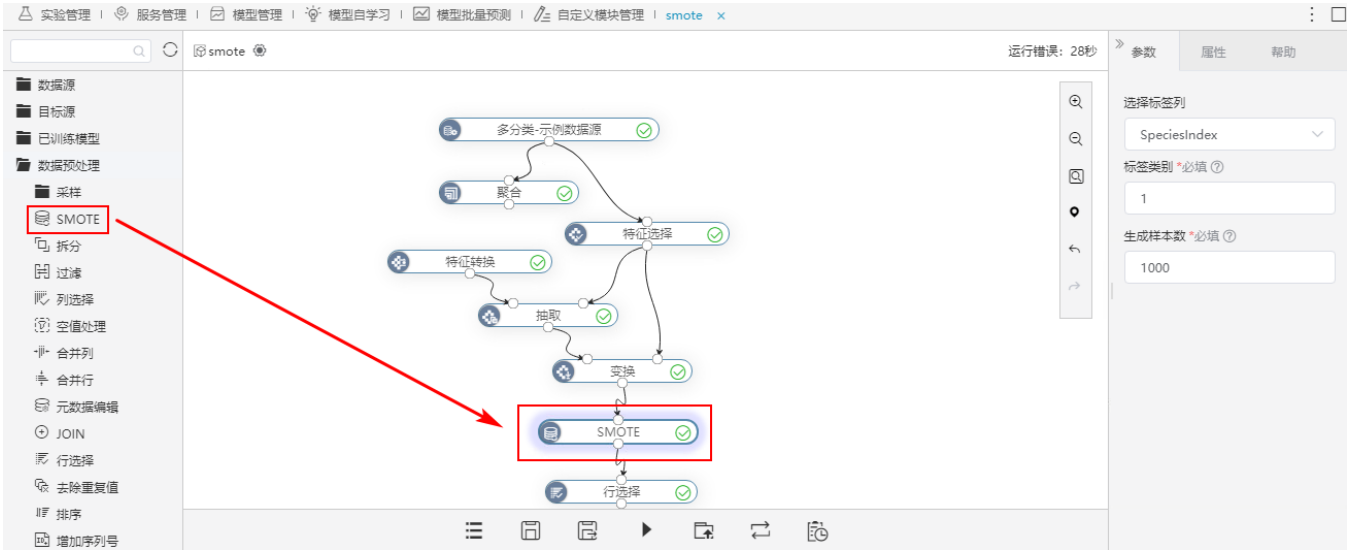
+【数据挖掘】新增SMOTE数据预处理方式

背景介绍

平时很多分类问题都会面对样本不均衡的问题，很多算法在这种情况下分类效果都不够理想。而SMOTE作为合成少数类过采样技术，是基于随机过采样算法的一种改进方案，可以用来解决类别不平衡问题。因此V10版本新增SMOTE节点，能够对少数类样本进行分析，并根据少数类样本人工合成新样本添加到数据集中。

功能简介

V10版本新增SMOTE节点，通过增加少数类样本的数量，使样本达到均衡。



参考文档

关于SMOTE节点，详情请参考 数据挖掘-SMOTE 。

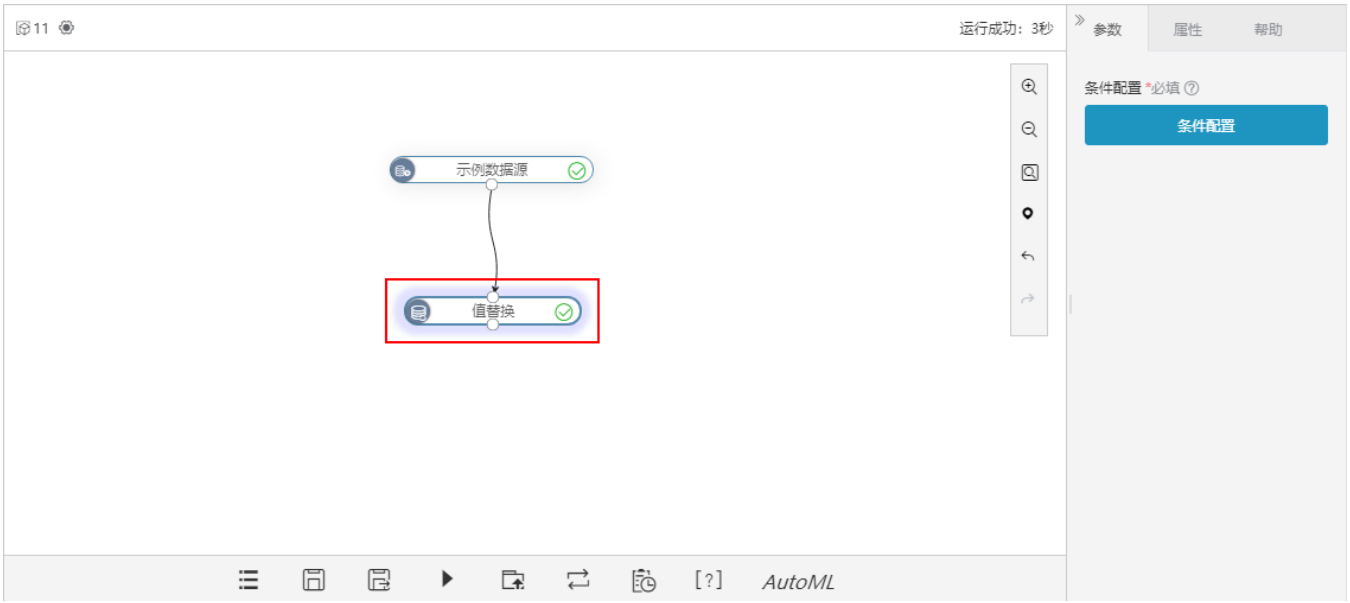
+【自助ETL/数据挖掘】数据预处理新增值替换节点

背景介绍

V10版本，在自助ETL和数据挖掘中新增值替换节点，可以对指定的数据进行替换，可以帮助用户替换掉数据中一些缺失、无效、错误的值。

功能简介

V10版本，在自助ETL和数据挖掘中新增值替换节点，可以对指定列进行值、字符串、正则替换。



参考文档

关于值替换功能，详情参考 数据挖掘-值替换。

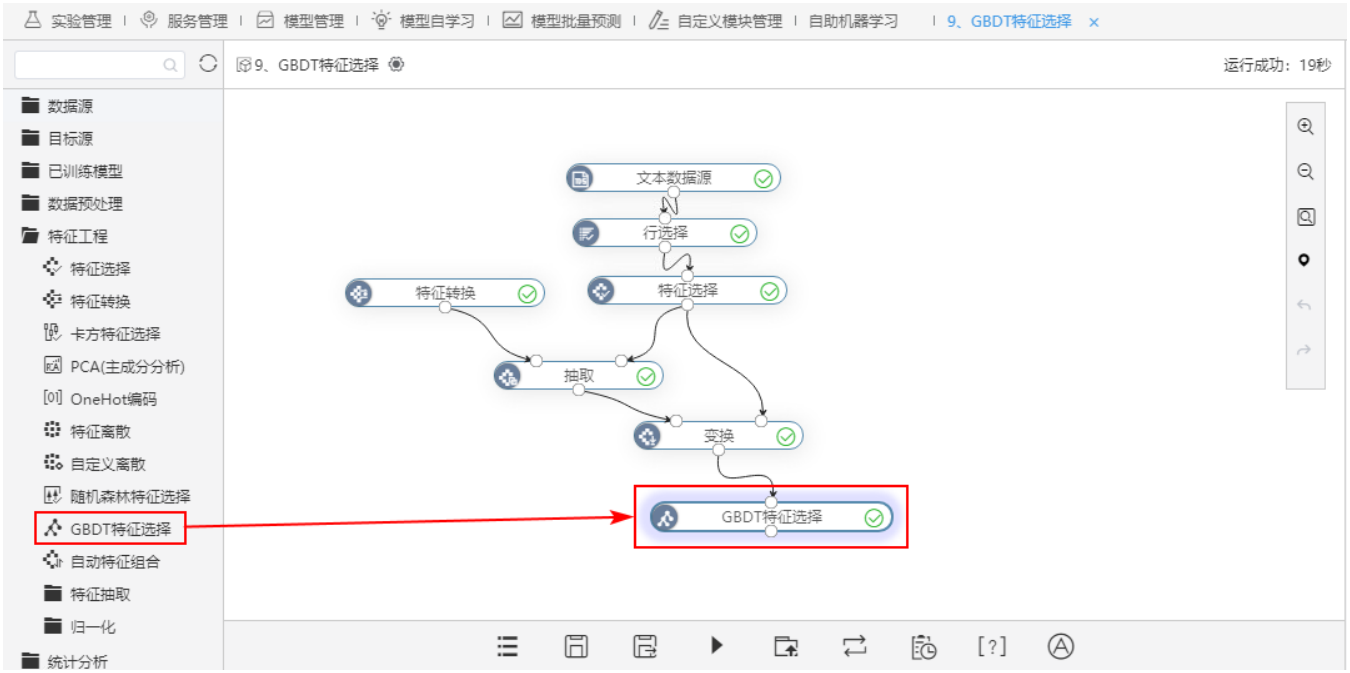
+【数据挖掘】特征工程新增GBDT特征选择节点

背景介绍

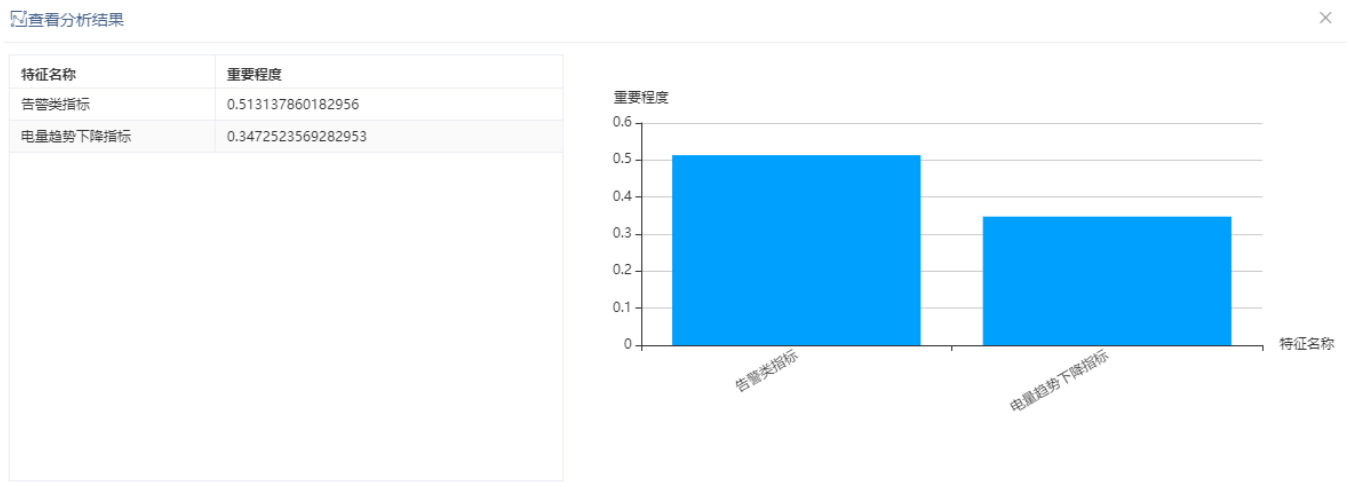
Smartbi现有的特征选择方法有卡方特征选择和随机森林特征选择，针对不同的数据情况有更丰富的特征选择方法及可对比性，V10版本新增GBDT特征选择节点。它的优势在于泛化能力强、模型输出后便于选择特征等。

功能简介

GBDT是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案。V10版本，左侧资源树特征过程节点下新增GBDT特征选择节点。



输出特征选择后的特征及其重要程度，以柱图展示如下：



详情参考

关于GBDT特征选择功能，详情参考 数据挖掘-GBDT特征选择。

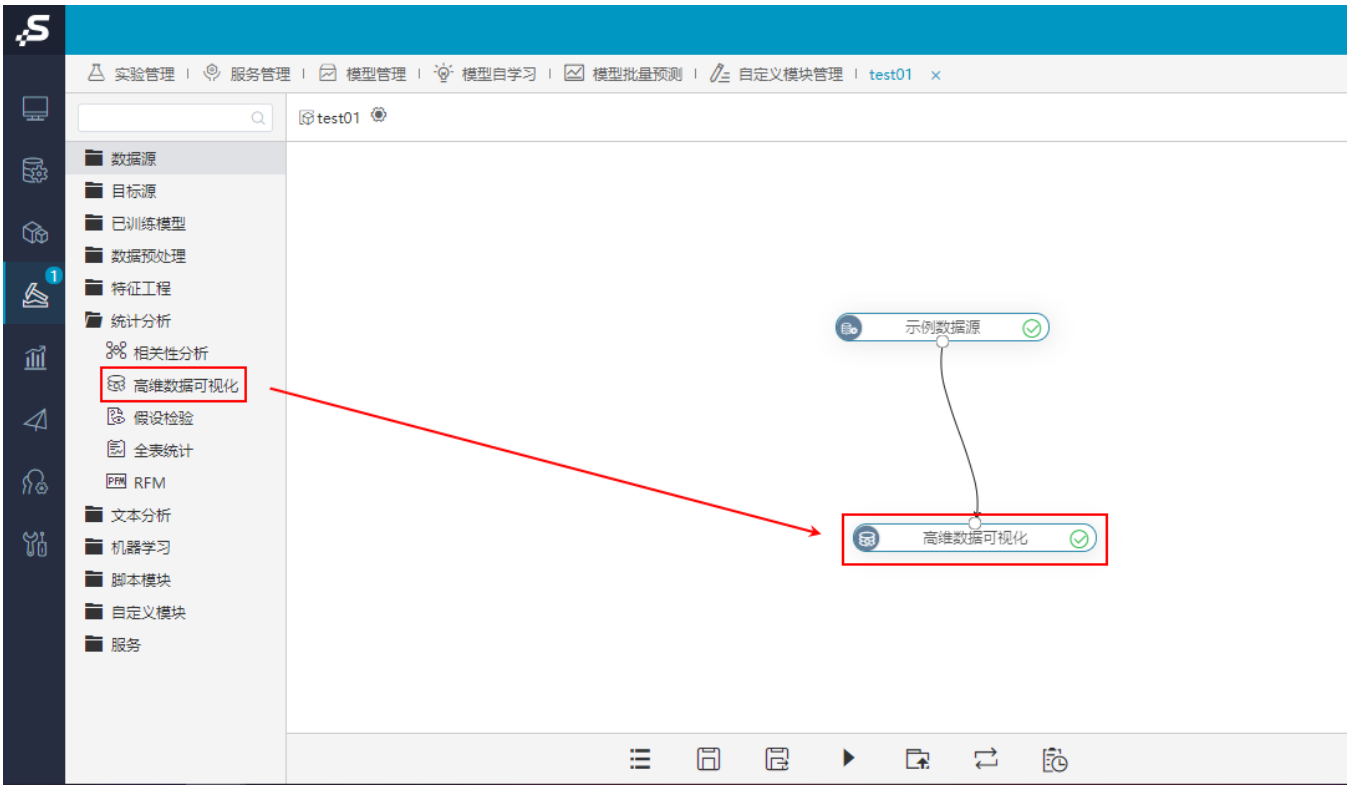
+【数据挖掘】统计分析支持高维数据可视化

背景介绍

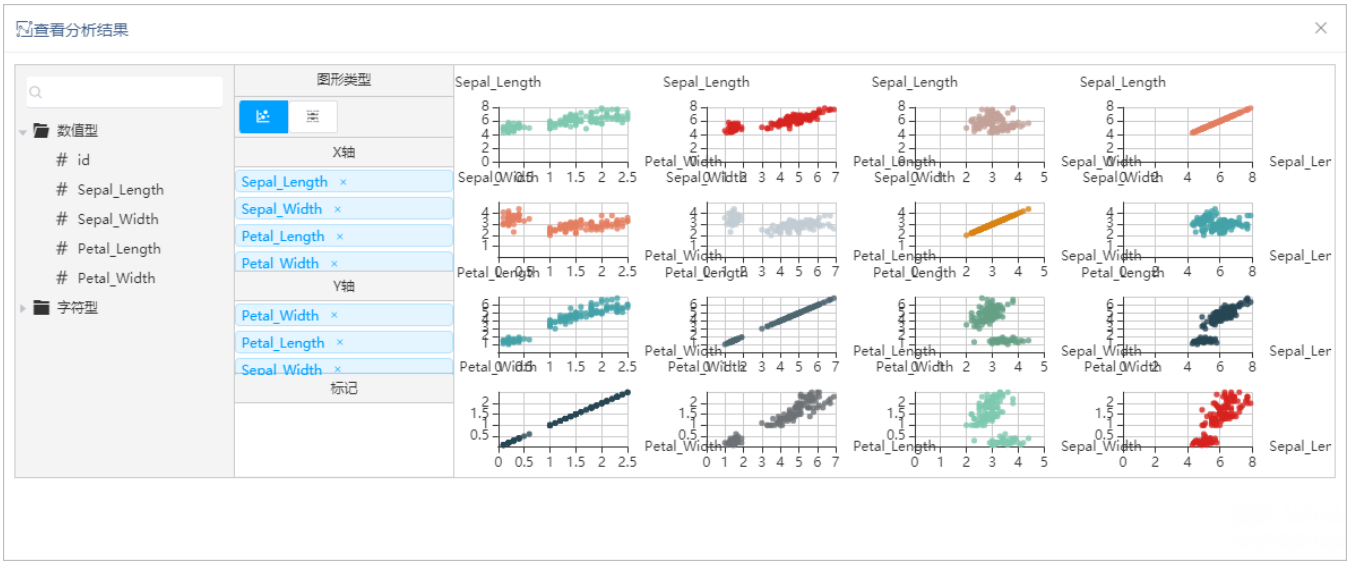
高维数据是指具有多个属性的数据，它在我们日常生活中十分常见，比如各种类型的多媒体数据、文档词频数据等等。面对这些高维数据，我们该如何展示各种属性之间的联系和发现它们之间的规律。其实在过去的数十年里，可视化领域已经产生了大量优秀的技术，如散点图矩阵、平行坐标图等，以帮助用户分析这类数据。

功能简介

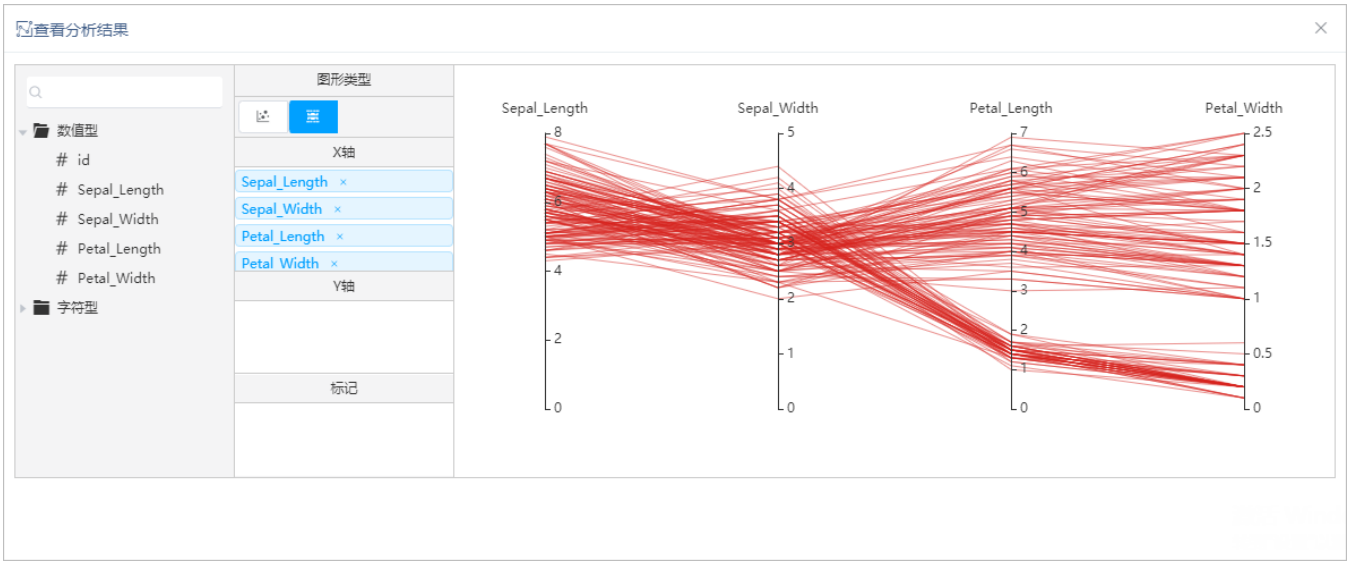
V10版本新增高维数据可视化节点，支持通过矩阵图和平行坐标图对高维数据进行可视化分析。



矩阵图效果：



平行坐标图效果：



详情参考

关于数据挖掘的高维数据可视化功能，详情请参考　数据挖掘-高维数据可视化。

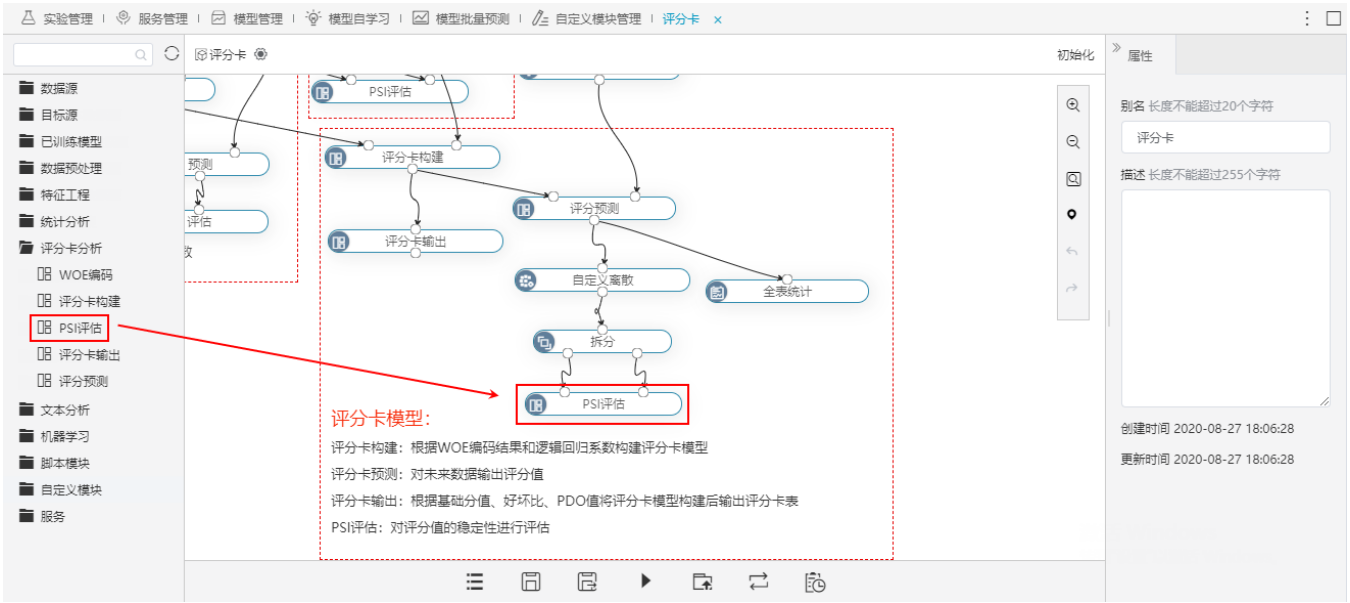
+【数据挖掘】评分卡分析新增PSI评估节点

背景介绍

在评分卡分析中，我们经常用到评分信用评级的分箱、数据转换模块、评分卡训练、评分卡预测等功能。支持评分卡模型应用后，还需对模型效果做评估，因此V10版本新增评分卡模型的PSI评估，用于对离散特征稳定性进行评估。

功能简介

V10版本新增PSI评估节点，用于对评分值的稳定性进行评估。



详情参考

关于PSI评估节点，详情请参考　数据挖掘-PSI评估。

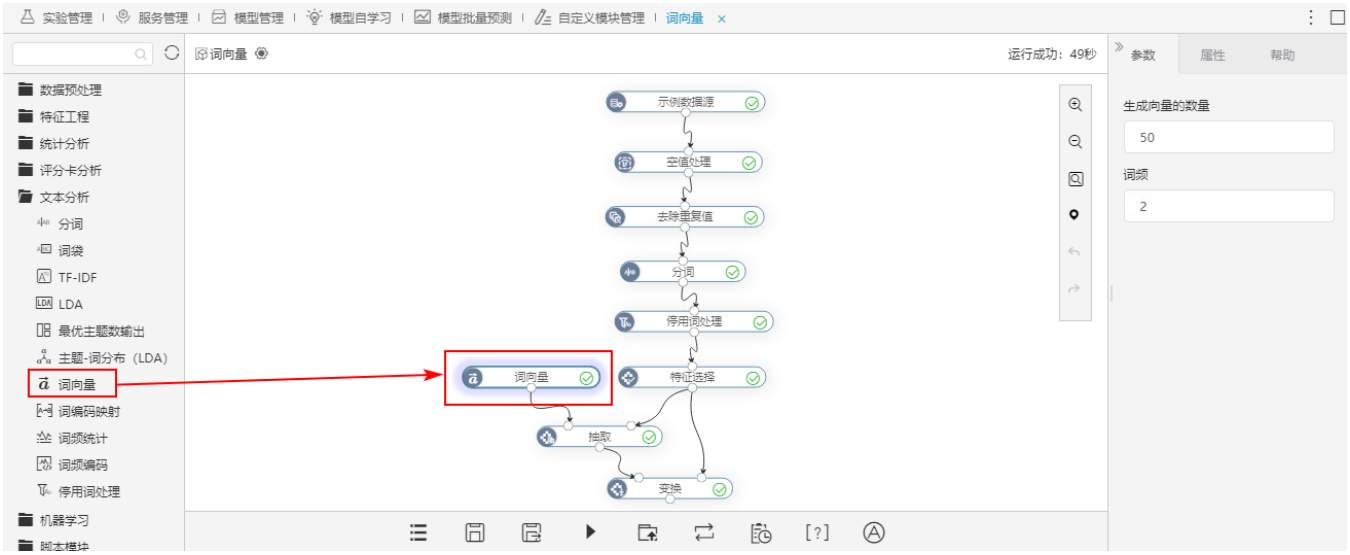
+【数据挖掘】文本分析增加词向量节点

背景介绍

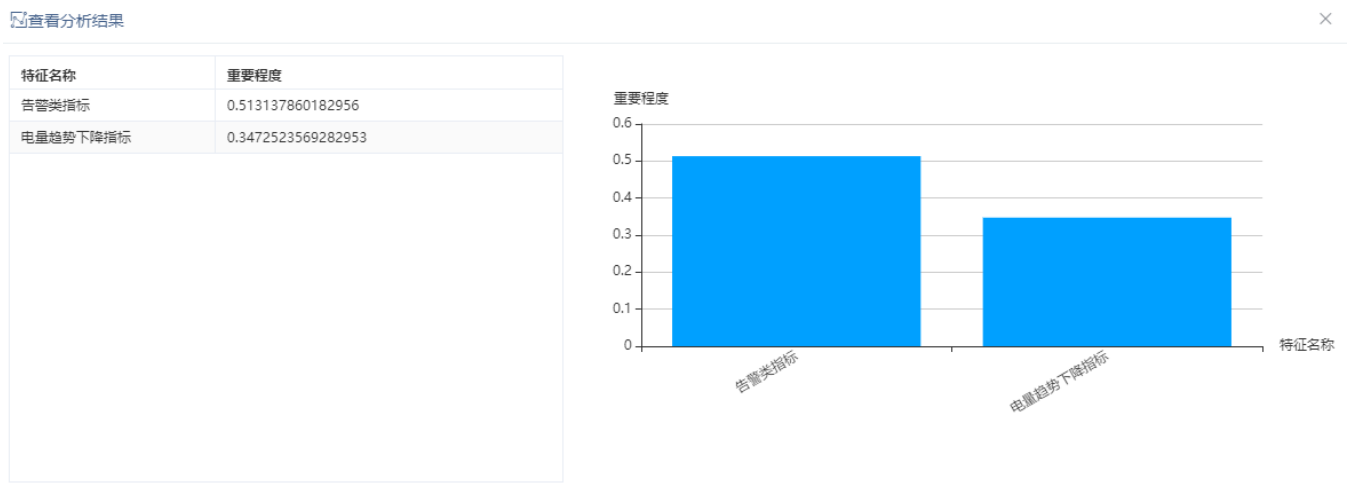
在文本分析中，我们会先采用词频编码，根据词频信息进行简单主题聚类或文本分类。但是这种方法忽略了词序信息，也无法判断出两个词语之间的关系。而Word2vec词向量可以很好地解决这个问题，它的思路是通过训练，将每个词都映射到一个较短的词向量上来。所有的这些词向量就构成了向量空间，进而可以用普通的统计学的方法来研究词与词之间的关系。

功能简介

词向量节点作为文本处理常用的特征工程手段、在情感分析、语义分析上可以用来增加模型准确性、计算相似性等功能。V10版本，左侧资源树文本分析节点下新增词向量节点。



在查看输出结果可以看到每个文本对应的词向量：



详情参考

关于词向量节点的功能，详情参考 数据挖掘-词向量 。

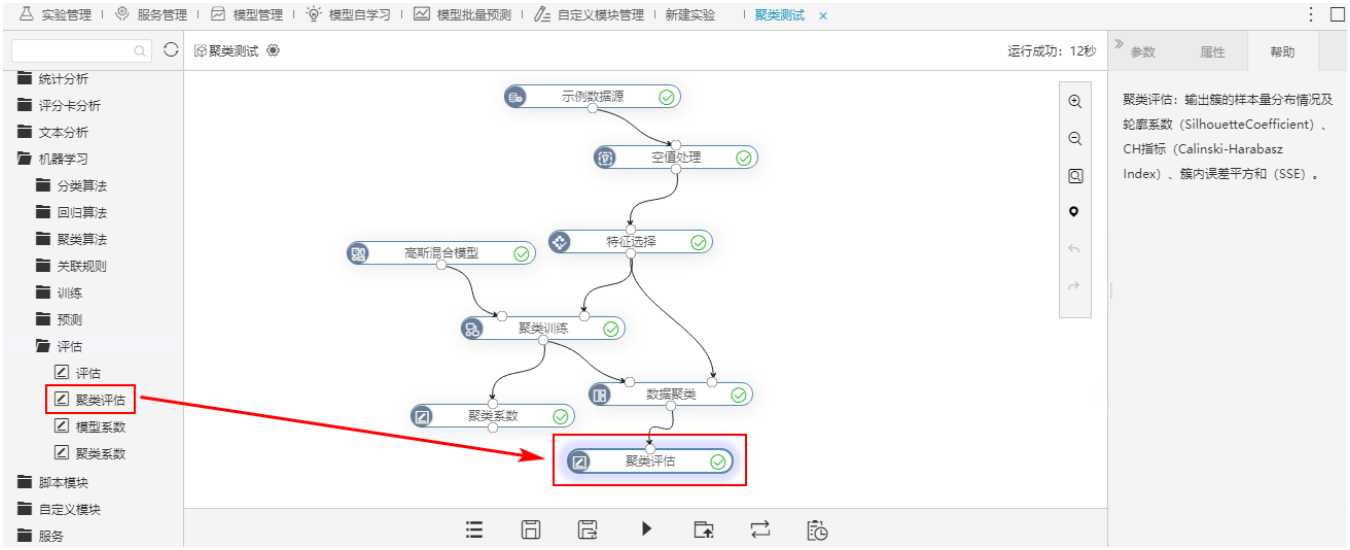
+【数据挖掘】新增聚类评估节点，用于呈现聚类算法常见评价指标值

背景介绍

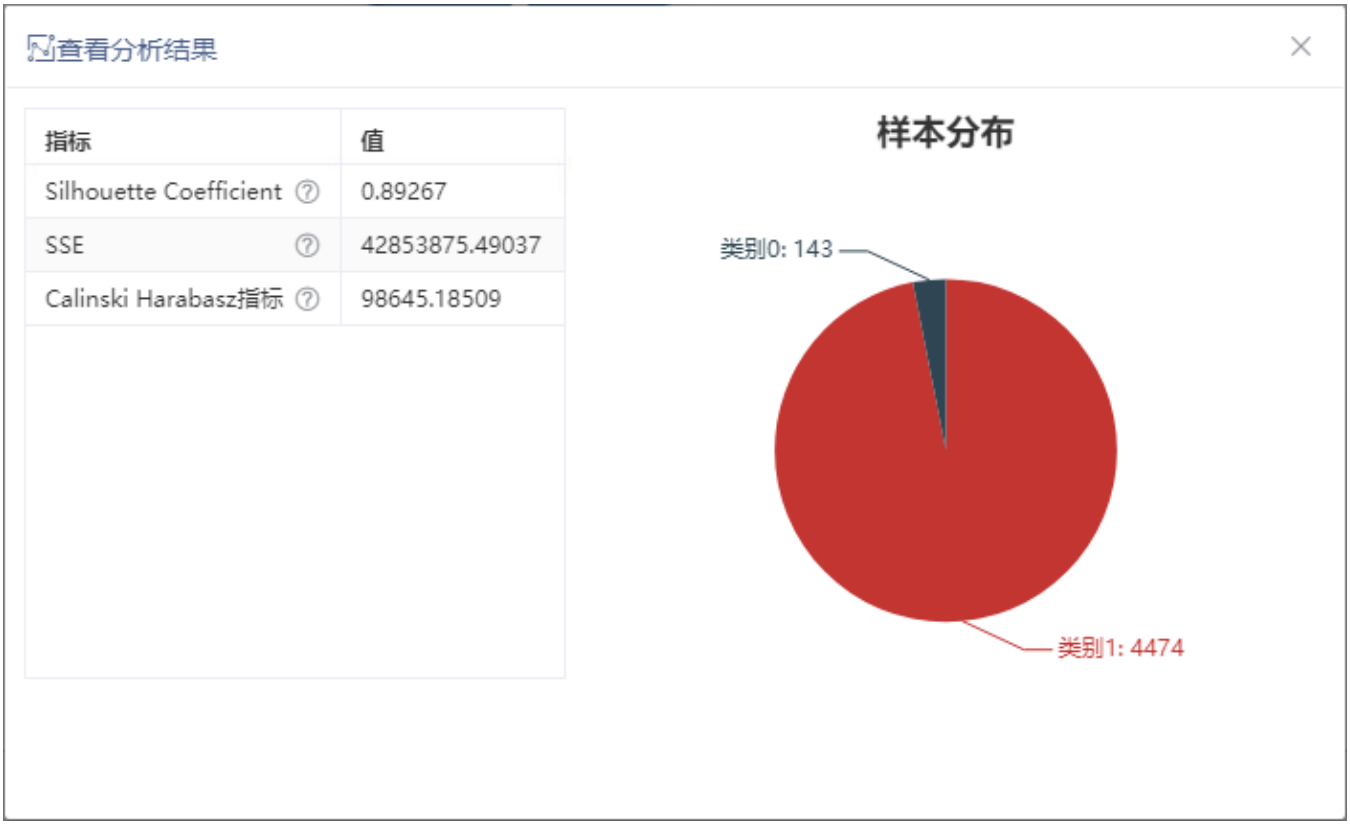
用户在做聚类时，往往无法直观的识别聚类结果的好坏，在数据质量不高的情况下，聚类的效果很不稳定，得出的结论也不容易让人信服。因此产品新增聚类评估节点，能够估计在数据集上进行聚类的可行性和被聚类方法产生的结果的质量，确保数据集聚类后的效果，使聚类结果更好的被应用到实际应用场景中。

功能简介

增加聚类评估节点，可以估计在数据集上进行聚类的可行性和被聚类方法产生的结果的质量。



分析结果包括对聚类算法的评估指标（轮廓系数、和方差、CH指标）和样本量分布情况，如图：



详情参考

关于聚类评估节点，详情请参考 数据挖掘-聚类评估 。

^【自助ETL/数据挖掘】关系数据源支持参数设置

背景介绍

以前的版本，用户想要在ETL或数据挖掘中使用参数切换数据，需要在执行之前人工干预数据流里的数据，操作繁琐也不够自动化。为了简化用户的操作，V10版本在自助ETL和数据挖掘模块中，关系数据源支持参数设置，用户可以通过改变参数查询条件值来改变数据，满足了用户不同的数据需求。

功能简介

在自助ETL和数据挖掘中，实验工具栏新增“参数设置”按钮，关系数据源新增SQL语句输入框，支持通过参数设置和在输入框中拼接SQL语句的方式来设置关系数据源的参数。

参数设置页面如下：

注意事项

在SQL输入框中，表达式的第一个字段使用的是表头真名。

详情参考

关于参数设置功能，详情请参考 数据挖掘-参数设置 。

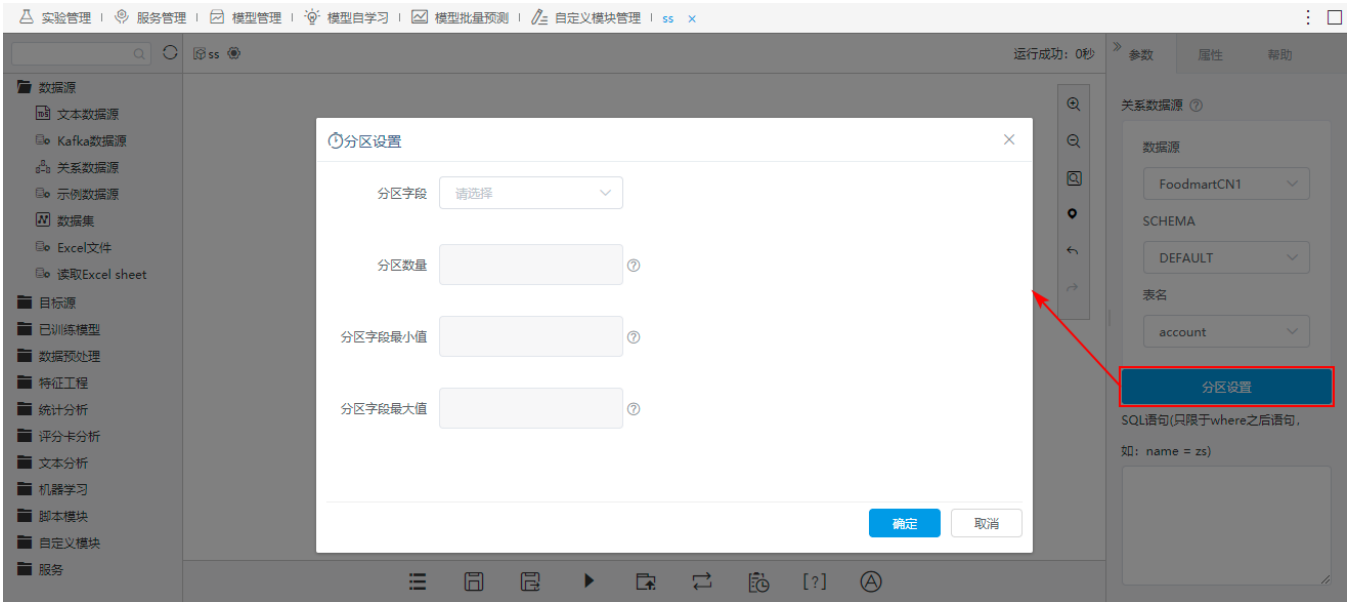
^【自助ETL/数据挖掘】关系数据源支持分区设置，提升数据抽取效率

背景介绍

数据抽取可以将源数据库的原始数据抽取到高速缓存库中，可以秒级获取大级别量的数据结果。为了进一步提升大数据量抽取性能，V10版本在自助ETL和数据挖掘中关系数据源支持分区读取数据，能够减轻系统压力，提升抽取效率。

功能简介

V10版本，关系数据源新增“分区设置”功能，可将数据分成几个区域后并行读取数据，提升数据抽取效率。



例如6KW的数据量提升抽取的效率如下：

数据量	不设置分区字段	设置分区字段
6KW	30分42秒	14分24秒

注意事项

Presto数据库暂不支持分区设置功能。

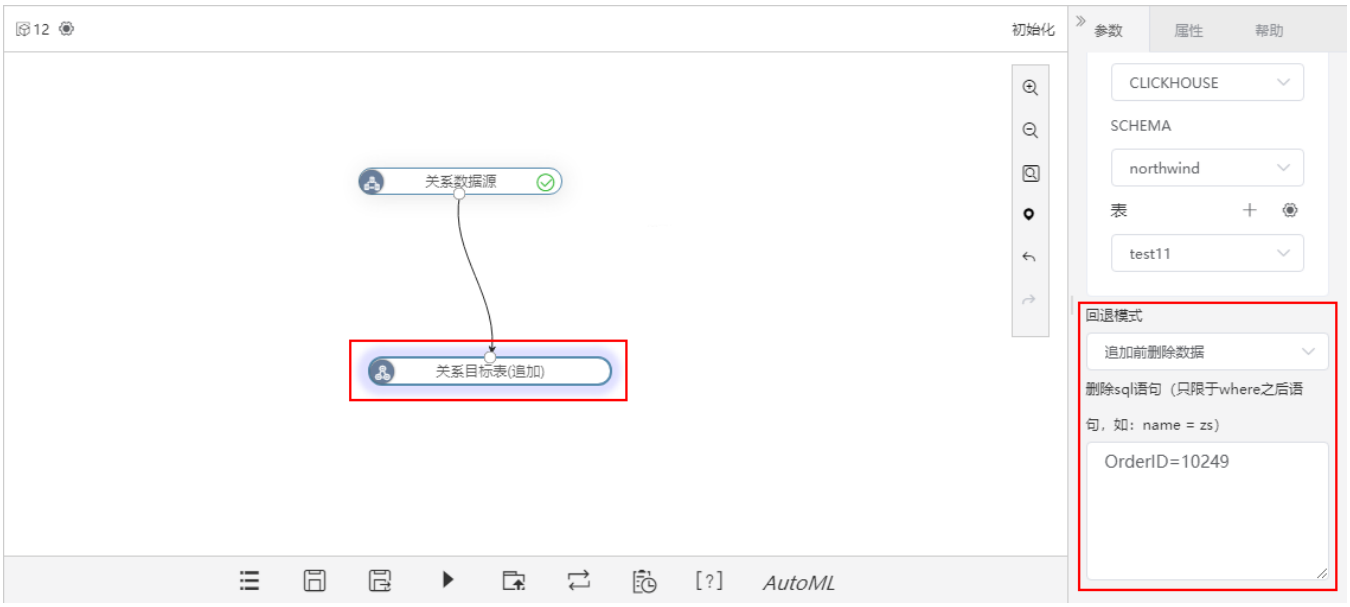
详情文档

关于关系数据源分区设置的功能，详情请参考 关系数据源 。

^【数据挖掘】关系目标表（追加）节点追加数据前支持删除表中数据

功能简介

V10版本，关系目标表（追加）节点追加数据前支持删除表中的数据，在回退模式中选择“追加前删除数据”并编写删除SQL语句，可以先删除表中部分或全部的数据，再将新数据追加到目标表中。



应用场景：用户在进行ETL调度时，发现某天调度的数据有问题，需要进行重跑（把之前已经入库的数据删除再插入），可以使用此功能可以先把入库的数据删除，再将新数据追加到目标表中。

注意事项

目前只有ClickHouse数据源（19.4.2.7版本及以上）支持此功能。

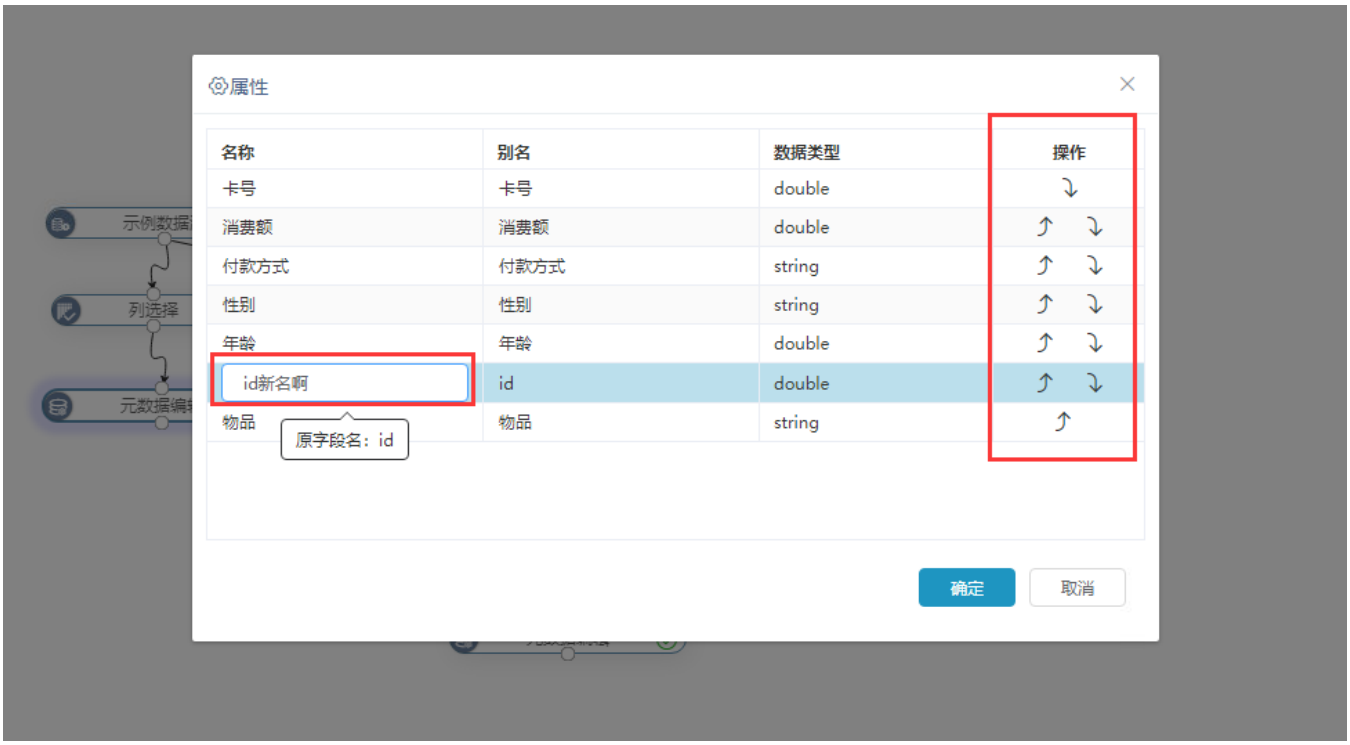
^【自助ETL/数据挖掘】元数据编辑支持修改原字段名及顺序

背景介绍

在实际场景中，Excel数据需要用到较多的数据处理操作，用户有修改元数据的原字段名和排序的一些需求。为了满足用户需求，V10版本元数据编辑节点支持修改原字段名及顺序，可以更全面地对数据进行处理，使数据更好地满足用户需求。

功能简介

在元数据编辑节点中，鼠标移动到名称列显示其原字段名，可修改数据的名称列。同时增加“操作”列，可对字段的顺序进行调整。



参考文档

关于元数据编辑的功能，详情请参考文档：数据挖掘-元数据编辑 。

^【数据挖掘】派生列、聚合、全表统计节点新增多个函数

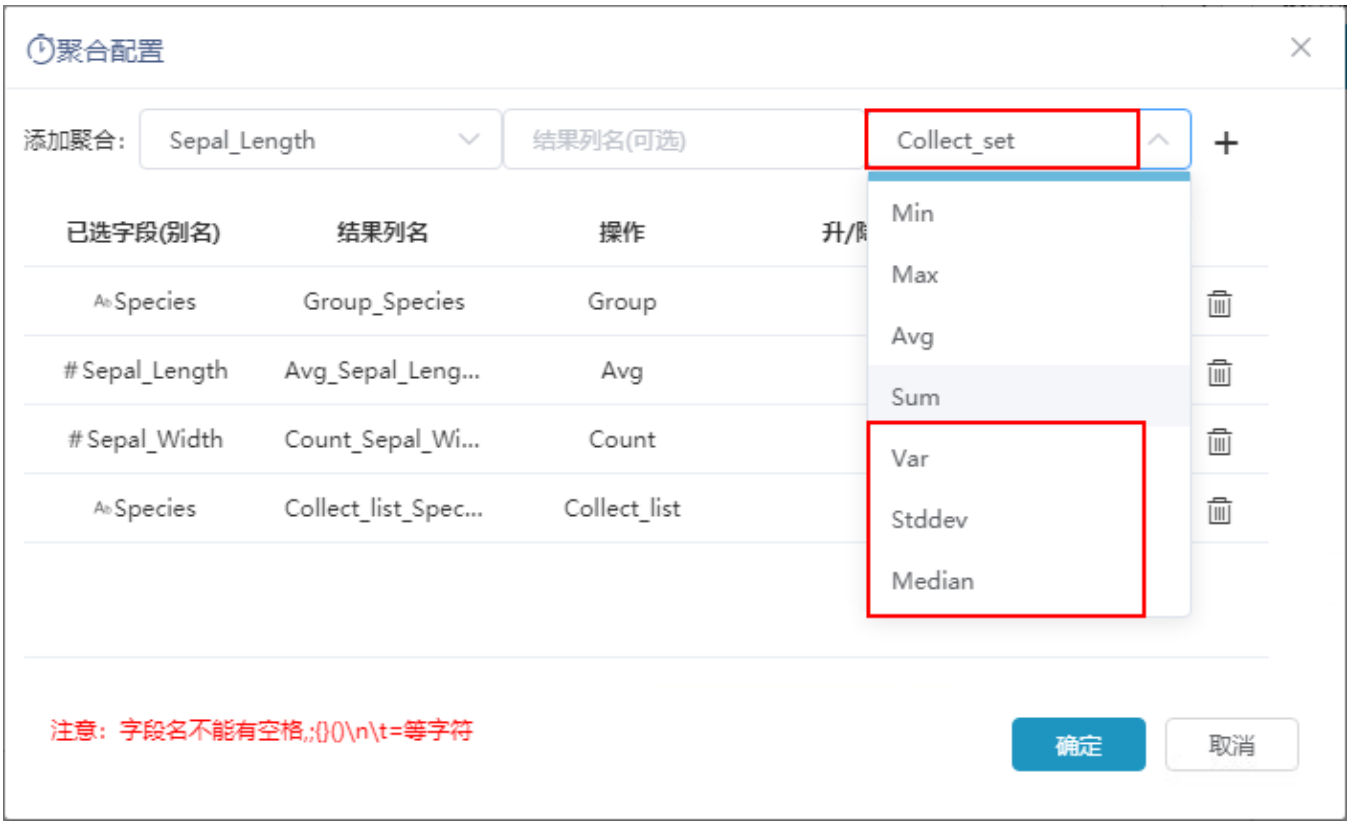
背景介绍

V10版本，派生列、聚合、全表统计节点新增多个函数，用于满足用户更多的需求，提升工作效率。

功能简介

1、派生列节点增加行最小值、行最大值函数。

2、聚合节点增加Collect_set 、方差、标准差、中位数等函数。



3、全表统计节点增加计算众数的方法。

参考文档

关于这些节点新增的函数功能，详情请参考文档：数据挖掘-派生列、数据挖掘-聚合、数据挖掘-全表统计 。

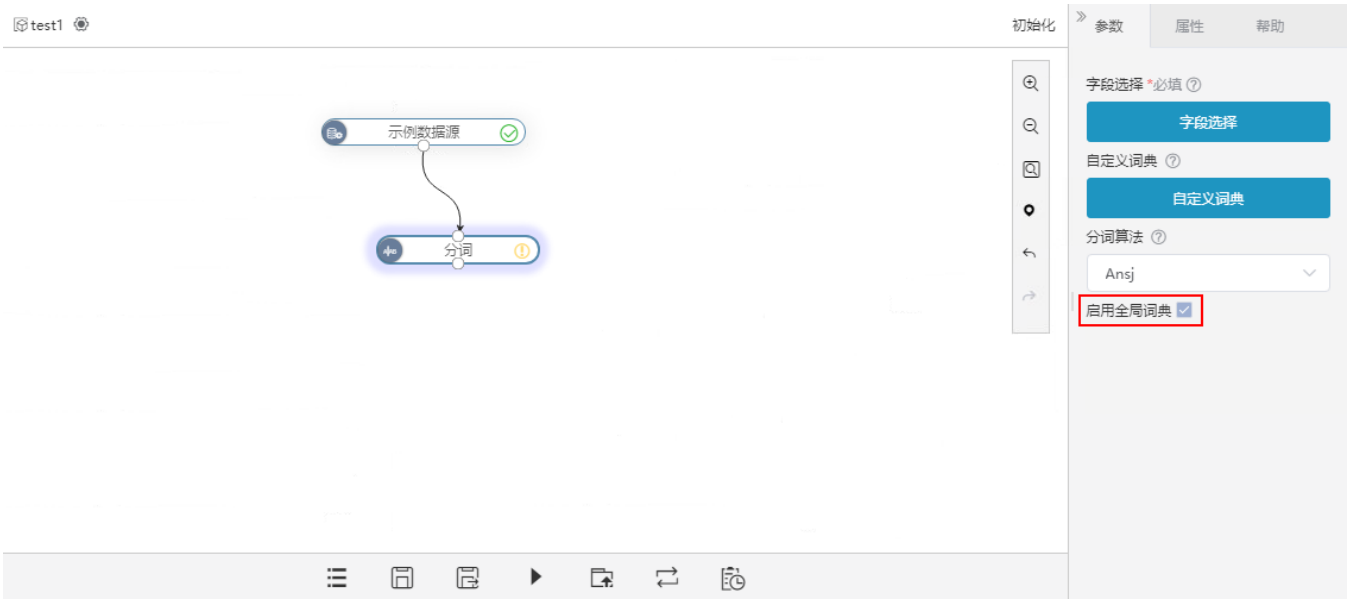
^【数据挖掘】分词节点新增自定义全局词典和分词算法

背景介绍

以前的版本，分词节点只在局部生效，无法同时满足用户多个节点的分词需求且效率较低。V10版本，新增自定义全局词典功能，用户上传自定义的分词可在全局使用，并新增了多个分词算法，可快速进行分词，提升分词效率，满足对分词效果要求高的各种场景。

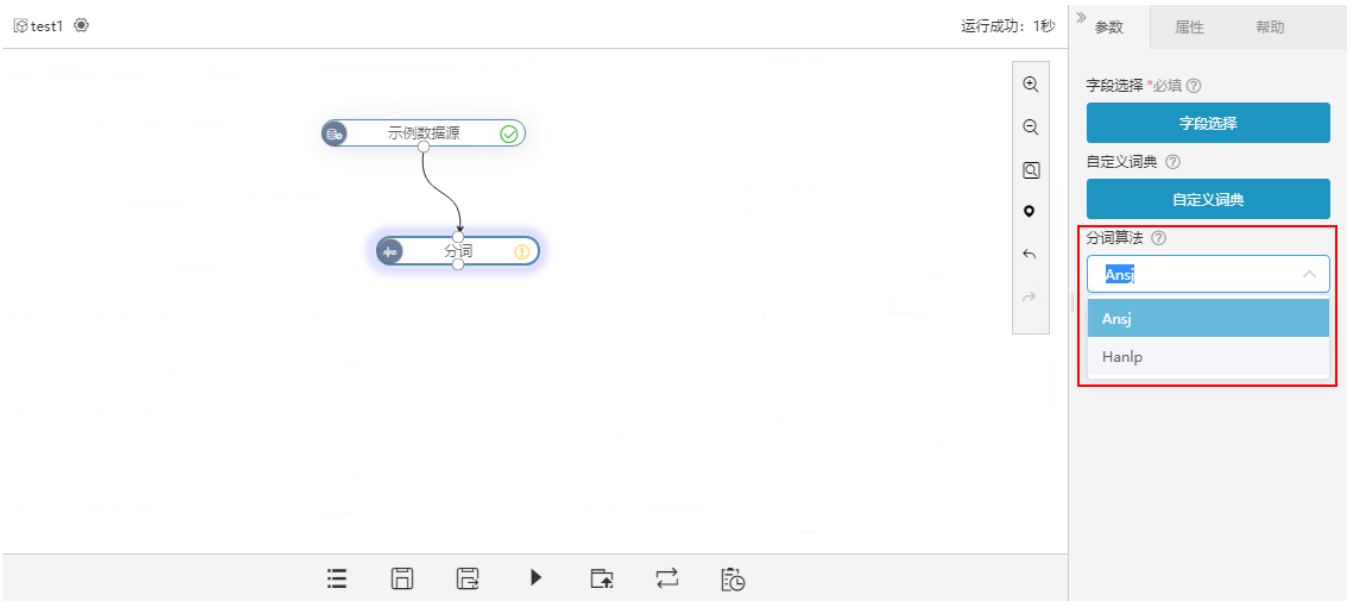
功能简介

1、分词节点新增“启用全局词典”设置项，可使用全局词典中的词辅助进行分词。



分词节点新增上传文件的方式上传自定义词典。

2、分词节点新增“分词算法”，可选择Ansj、Hanlp算法。



详情参考

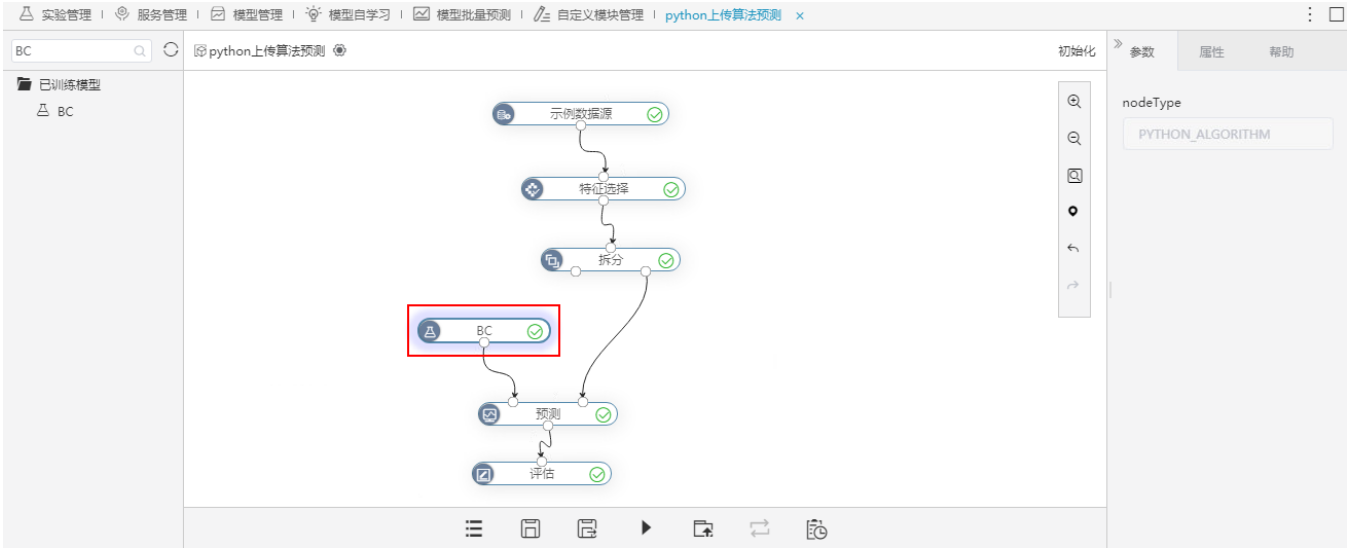
关于全局词典和分词算法，详情请参考 数据挖掘-分词 。

^【数据挖掘】完善Python算法节点功能

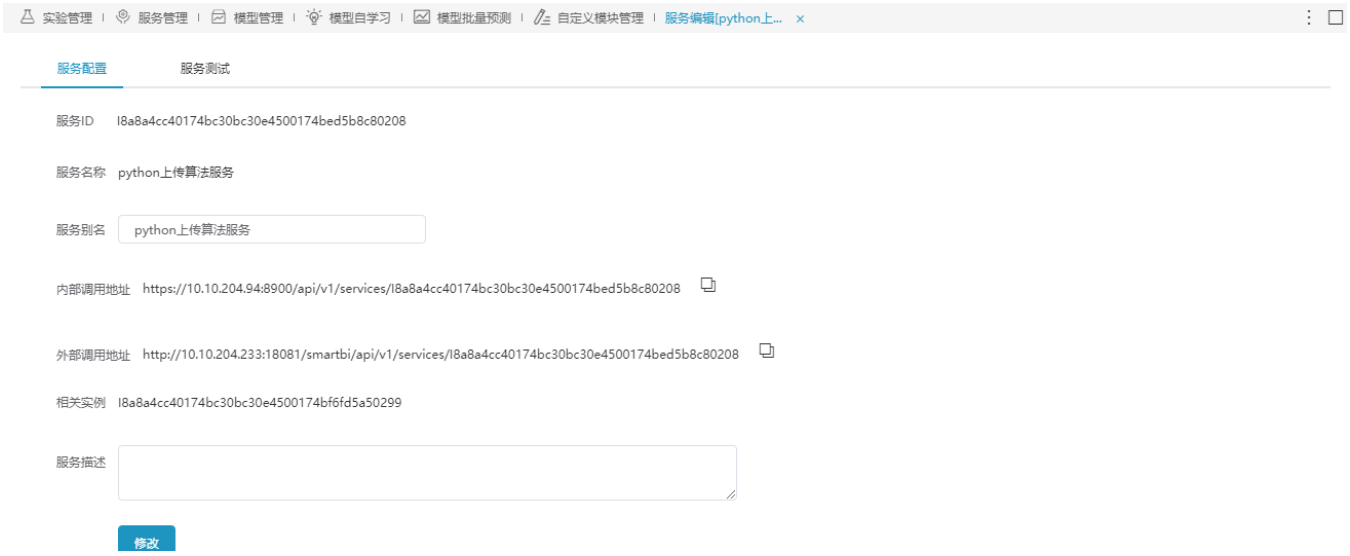
功能简介

V10版本，我们完善了上传的Python算法节点功能，能够在产品中进行模型训练、模型保存、模型预测、模型评估、服务等。

示例1：上传的Python算法节点进行模型预测、评估。



示例2：上传的Python算法节点进行部署服务。



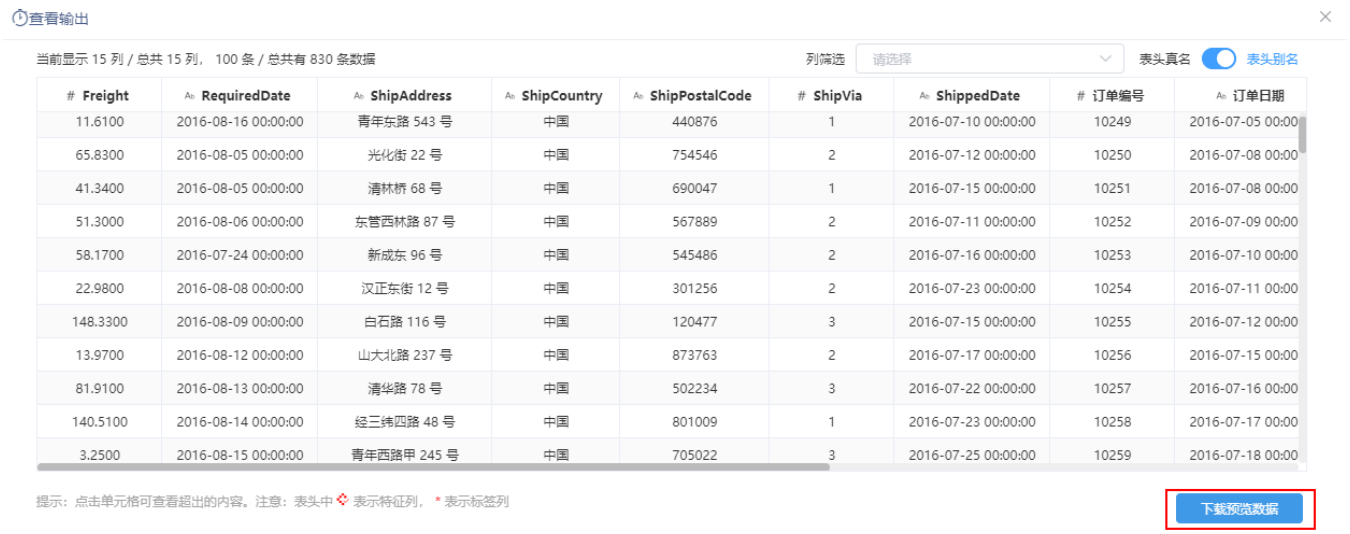
^【数据挖掘】查看输出支持预览数据导出到本地

背景介绍

在挖掘实验过程中，对每一个执行完的节点资源我们都可以预览该节点的数据，如果可以将预览数据导出到本地，这将便于用户进行后续的处理或分析。

功能简介

V10版本支持预览数据导出到本地，在查看输出窗口新增“下载预览数据”选项。



注意事项

此处会把预览的数据以csv文件的方式下载到本地，不会下载全量数据，数据量最多100条。

^【自助ETL/数据挖掘】查看输出增加列筛选项

背景介绍

在自助ETL或数据挖掘实验中，对每一个执行完的节点查看输出数据时，能显示的数据量有限；V10版本中，增加对字段进行列筛选过滤的功能，方便用户查验数据。

功能简介

在节点“查看输出”页面新增列筛选功能，对输出数据进行筛选，方便用户查看。

对于列筛选后的数据仅限于查看，下载预览数据仍是对筛选前的数据进行下载。

^【数据挖掘】节点输出字段支持排序

功能简介

V10版本，节点输出字段的顺序按照选择字段的先后顺序排序。

输出的字段顺序如图：

注意事项

1、WOE编码、异常值处理节点不支持排序。

2、有些没有数据输出的节点，在节点设置时会显示选择节点的顺序，但输出时仍按照原始顺序排序，如特特征选择节点。

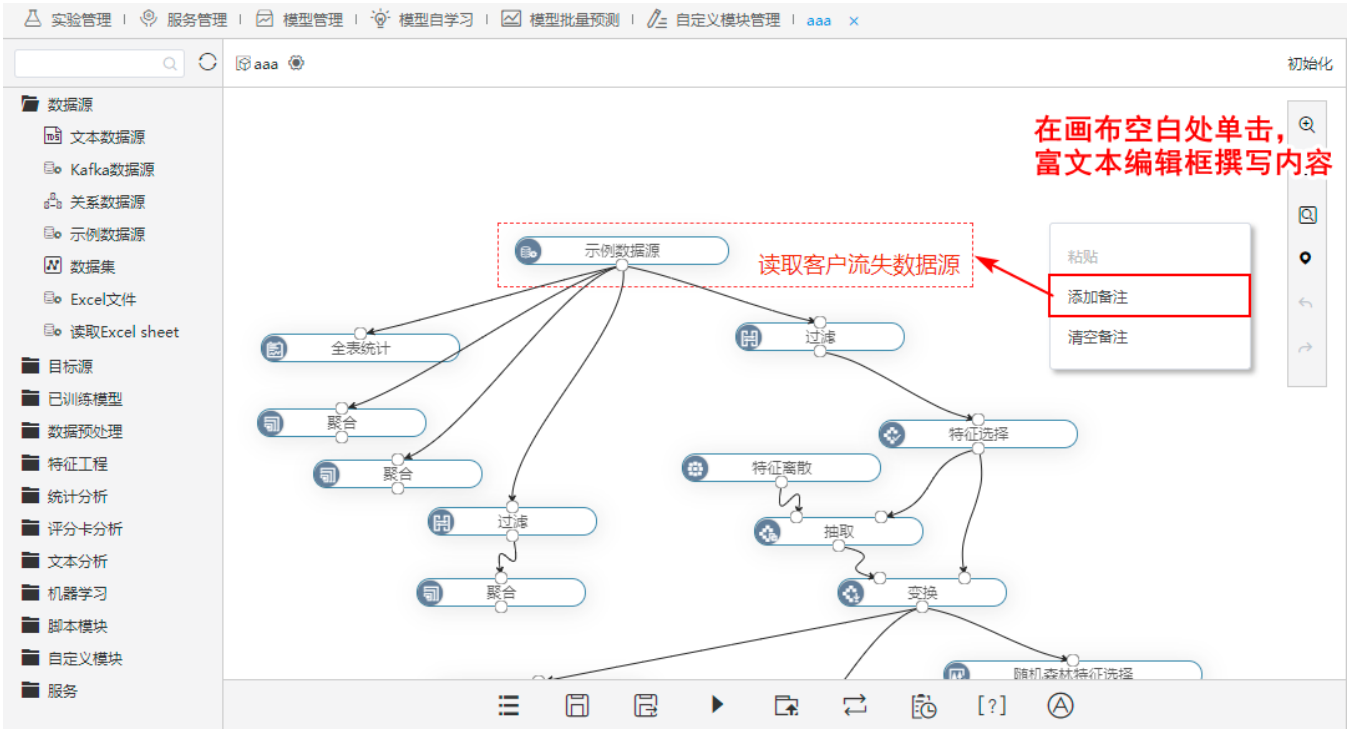
^【数据挖掘】增强整个页面的操作

背景介绍

机器学习实验往往牵涉多个节点，各节点之间关系也较为复杂，更或者自定义的算法节点只有实验构建者才明白其中的含义；同时在实验构建过程中，可能出现节点复用的情况。因此在V10版本中，在实验和节点增加备注功能、节点增加复制功能，便于实验的交流和提高实验的构建速度。

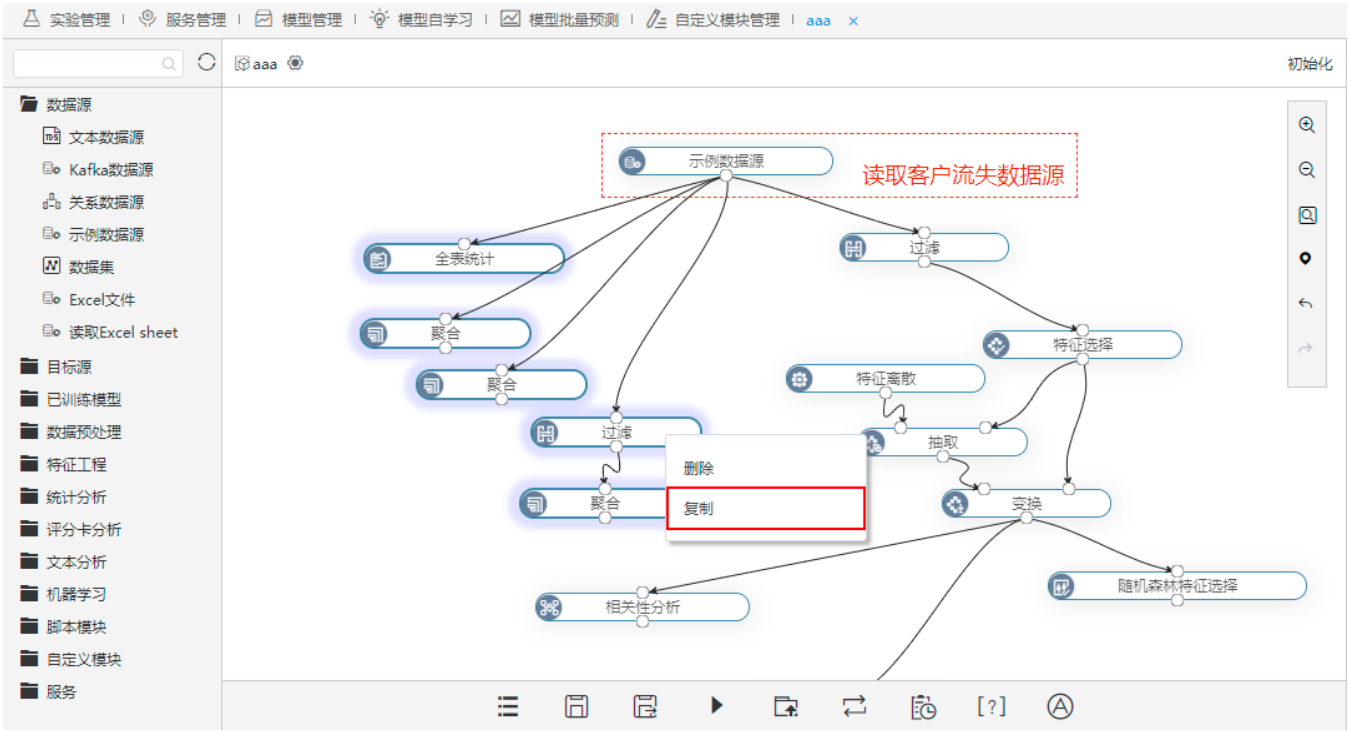
功能简介

在画布空白处单击右键，选择“添加备注”，会弹出富文本编辑框，可以添加对实验背景的介绍等内容。



选中需要复制的节点，单击右键，出现‘复制’，也可以同时选中多个节点：

- 拖动鼠标覆盖需要选择的节点，箭头滑过的矩形区域的节点都被选中（缩放状态下不支持框选），选中后可一起拖拽移动；
- 按住Crtrl键，鼠标逐个单击需要复制的节点，选中后可一起拖拽移动。



详情参考

关于节点的备注和复制功能，详情请参考　数据挖掘-实验界面介绍　。

^【自助ETL/数据挖掘】支持缓存节点数据，减少执行实验等待时间

功能简介

V10版本，数据挖掘新增“缓存节点数据”设置项（安装部署过Hadoop才生效），支持缓存执行过的节点的数据，下一次执行可直接执行当前配置好及其之后的节点，减少等待时间，提高工作效率。

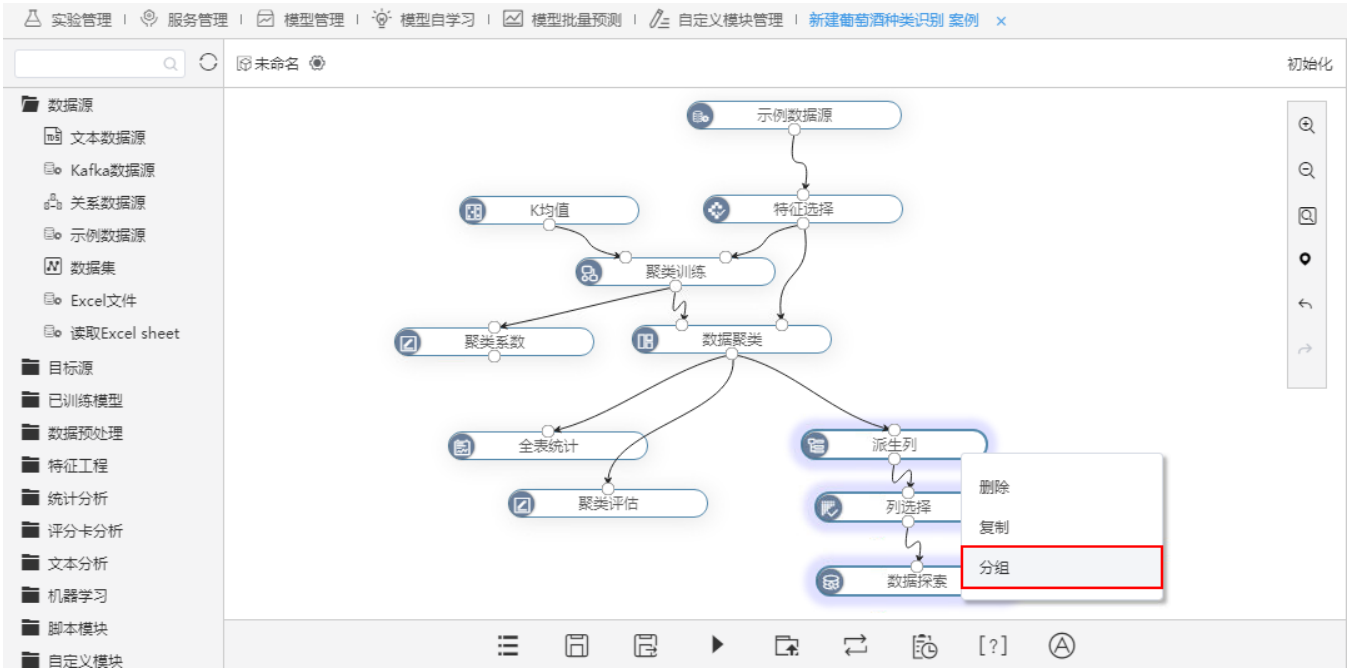
参考文档

关于缓存节点数据功能，详情请参考　缓存节点数据　。

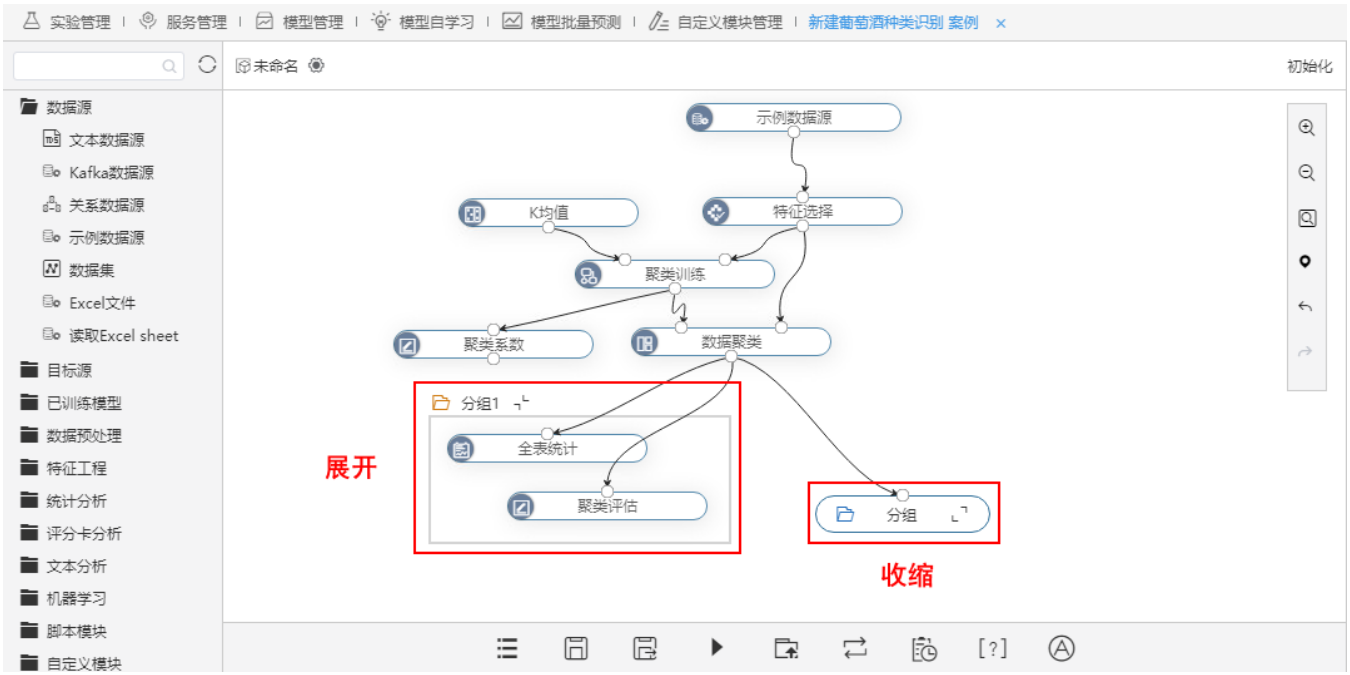
^【自助ETL/数据挖掘】支持多节点分组收缩和展开

功能简介

在自助ETL和数据挖掘实验中，支持选择多个节点合并为一组，以便节点较多的实验归类和移动节点。



同组的节点可收缩或展开：



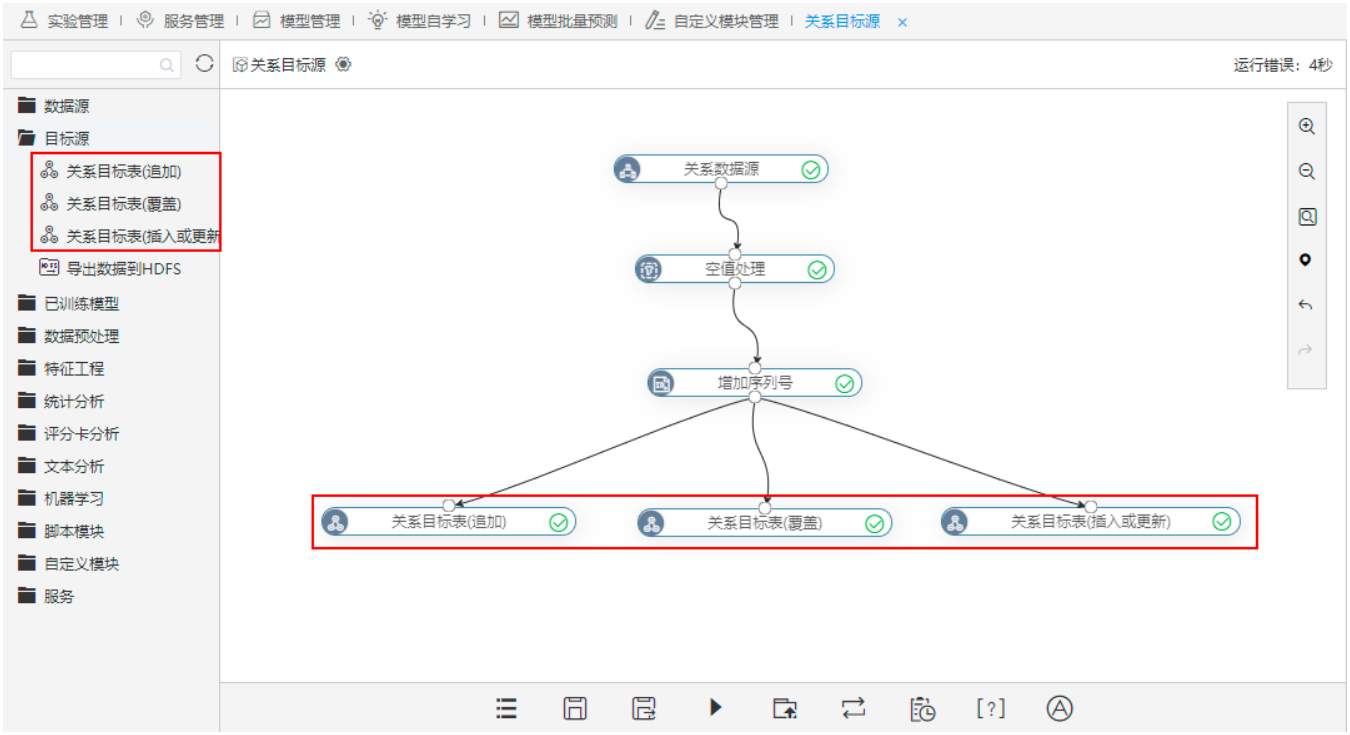
〈【自助ETL/数据挖掘】关系目标源拆分为追加、覆盖、插入或更新数据节点

背景介绍

以前的版本，用户在数据挖掘和自助ETL中，只能通过追加的方式导出处理和分析后的数据，方式单一。为了满足用户需求，V10版本在自助ETL和数据挖掘中，可以使用追加、覆盖、插入或更新的方式导出数据，以便用户能够针对不同的情况选择不同的方式插入数据。

功能简介

在自助ETL和数据挖掘中，关系目标源分为关系目标表（追加）、关系目标表（覆盖）、关系目标表（插入或更新），用户可以通过这三种方式将数据导出到目标库中。



详情参考

关于关系目标表的导出功能，详情请参考 目标源 。

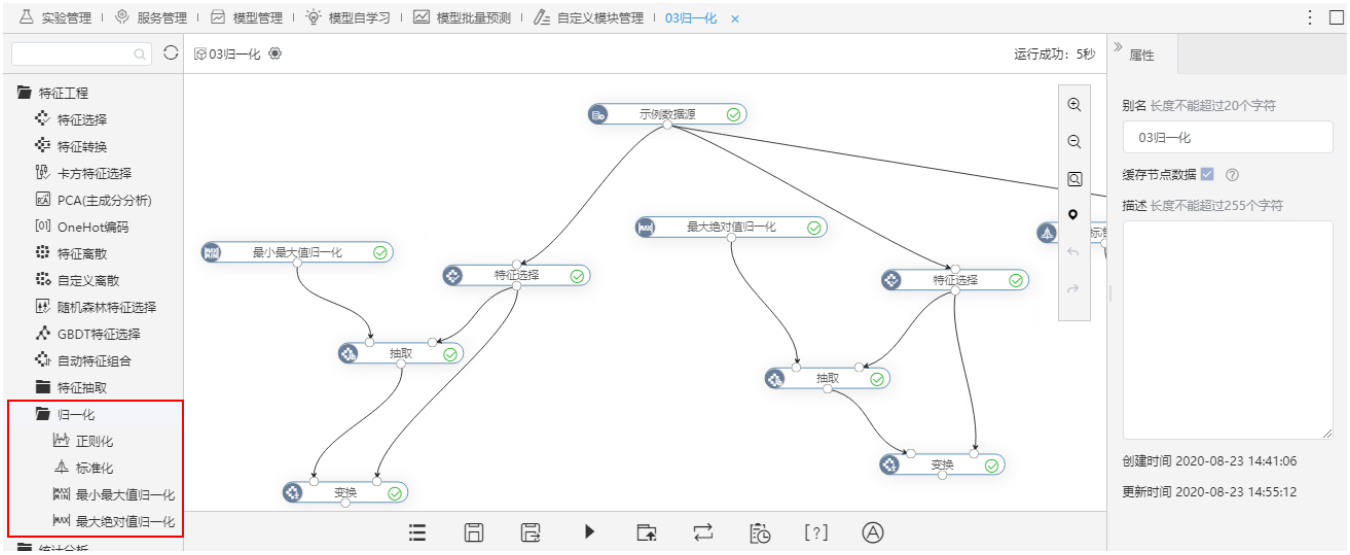
〈【数据挖掘】拆分归一化算法为多个节点

背景介绍

数据预处理在众多机器学习算法中都起着重要作用，实际情况中，将数据做归一化处理，消除量纲可以加速优化过程，使模型更好、更快的达到收敛。而在此之前Smartbi的归一化算法是封在其他算法当中，因此为了满足灵活性的需要，V10版本拆分归一化算法为多个节点。

功能简介

Smartbi的归一化算法有四种，分别为：正则化、标准化、最小最大值归一化、最大绝对值归一化。



详情参考

关于归一化节点功能，详情参考 数据挖掘-归一化 。