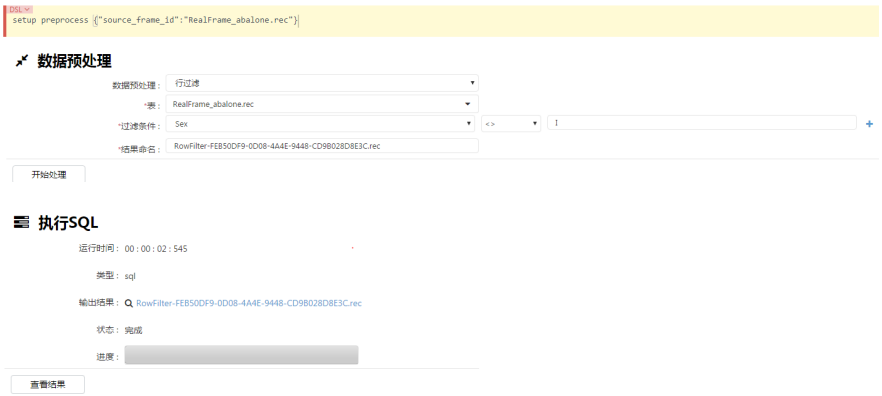
ADE数据预处理算子不同处理方法实例设计

王乐A 2018-03-14 发表

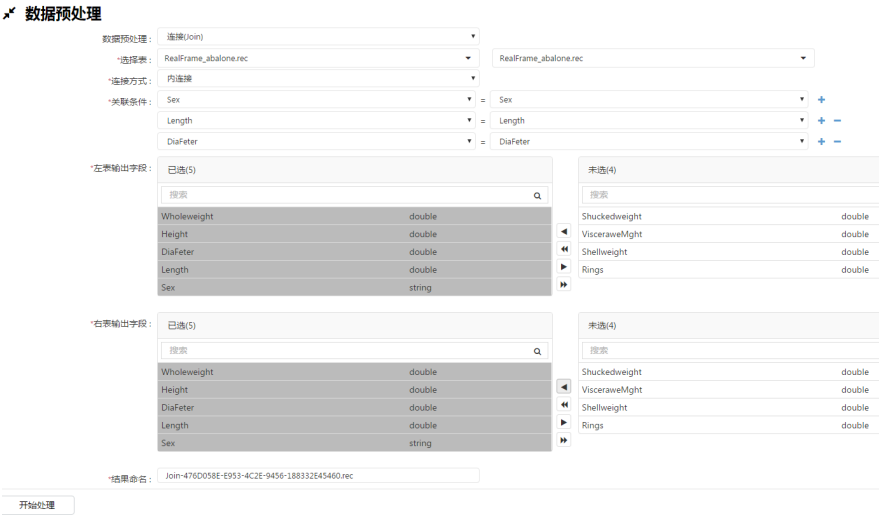
针对数据分割、数据标准化、缺失值填充、主成分分析、哑编码、join、union、行过滤、列过滤、数据均衡、增加序列号等预处理或者特征工程，做统一实例分析

通过实际项目测试预处理算子做此总结，注意事项：Word2vec、TF-IDF、LDA和文本处理有关，目前只能针对结构化数据中其中某些列中含有文本进行处理分析，不能对文档分析

1、 行处理，针对一些存在不合理的数据进行过滤：

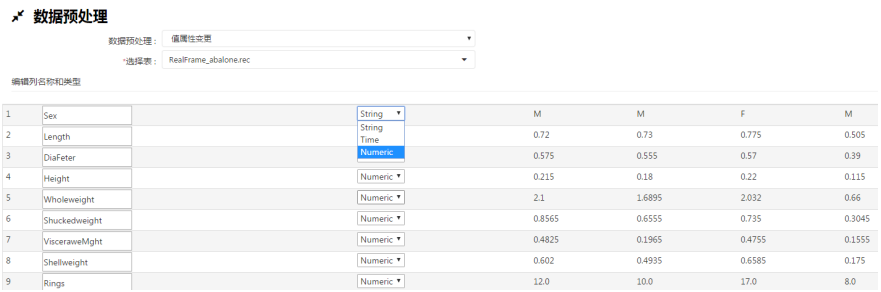


2、 Join，针对多表关联



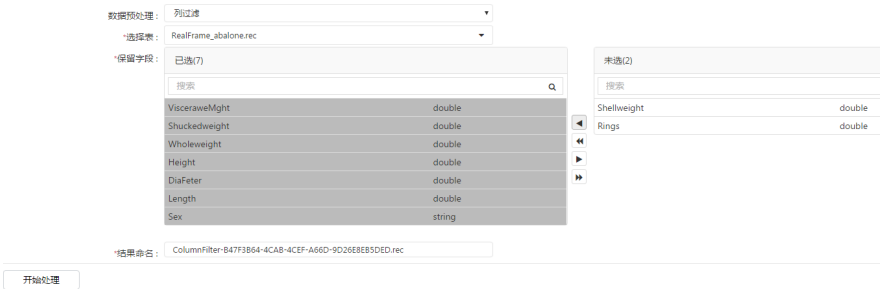
3、数据均衡，针对样本中不同类数据比例不一致适用该方法，通过采样法把不平衡的数据修正为平衡的数据，分为过采样和欠采样，其中欠采样法主要是对大类进行处理，过采样法针对小类进行处理该方法也被称作升采样(Upsampling)，优势是没有任何信息损失，但很有可能导致过拟合，SMOTE法是一种人工数据合成的过采样技术，目前平台只有过采样方法：Upsampling和SMOTE，因为SMOTE需要合适的样本选择近邻个数，所以本次用Upsampling来验证：

4、值属性变换，针对数据类型不符合的情况：



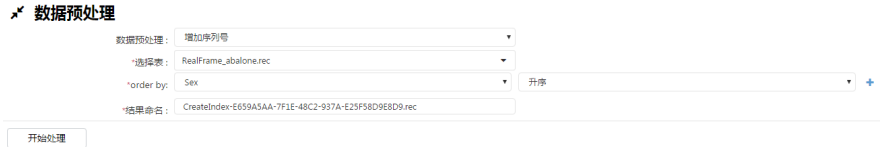
5、列过滤，针对数据集中一些无关紧要的列进行过滤，只保留部分列进行分析建模:

数据预处理



6、 增加序列号，对数据集某列排序，查看数据集:

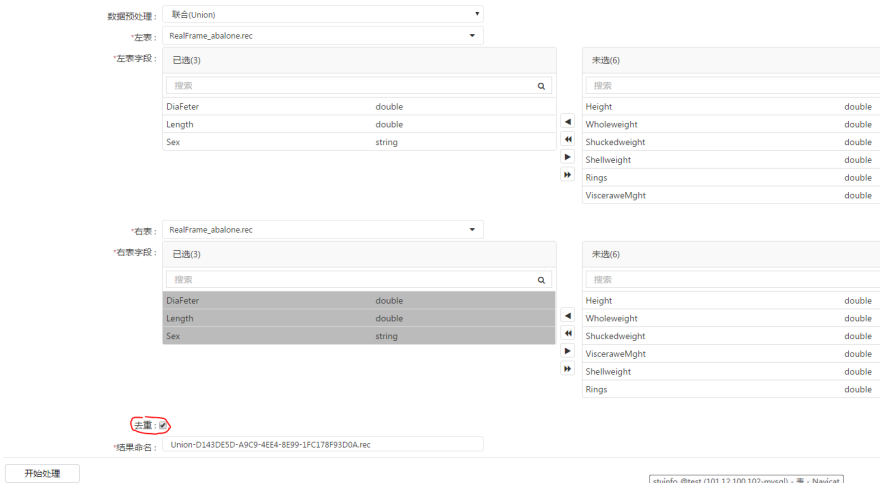
数据预处理



row_num	Sex	Length	DiaFeter	Height	Wholeweight	Shuckedweight
1.0	F	0.775	0.57	0.22	2.032	0.735
2.0	F	0.53	0.425	0.13	0.7455	0.2995
3.0	F	0.665	0.535	0.225	2.1835	0.7535
4.0	F	0.62	0.49	0.17	1.2105	0.5185
5.0	F	0.525	0.405	0.13	0.7185	0.3265
6.0	F	0.425	0.325	0.1	0.3295	0.1365
7.0	F	0.385	0.295	0.095	0.335	0.147
8.0	F	0.47	0.375	0.115	0.4265	0.1685
9.0	F	0.62	0.51	0.175	1.1505	0.4375
10.0	F	0.595	0.47	0.15	0.8915	0.359

7、Unlon，针对两个分散的数据集，进行合并，可选择是否去重，注意列数量和数据类型需要一致:

数据预处理



通过项目实际测试总结了针对我们平台的通用的算子预处理方法，实际应用场景中需要根据不同的数据集选择不同的处理方法，此测试例只为验证支持通过，可支持自定义扩展Word2vec、TF-IDF、LDA和文本处理有关，目前只能针对结构化数据中其中某些列中含有文本进行处理分析，不能对文档分析

。