

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：

数据预处理：行过滤
表：RealFrame_abalone.rec
过滤条件：Sex
结果命名：RowFilter-FE850DF9-0D08-4A4E-9448-CD98028D8E3C.rec

开始处理

执行SQL
运行时间：00:00:02.545
类型：sql
输出结果：RowFilter-FE850DF9-0D08-4A4E-9448-CD98028D8E3C.rec
状态：完成
进度：[Progress bar]

查看结果

2、Join，针对多表关联

数据预处理：连接(join)
选择表：RealFrame_abalone.rec
连接方式：内连接
关联条件：Sex = Sex, Length = Length, DiaFeter = DiaFeter
左表输出字段：已选(5)
右表输出字段：已选(5)
结果命名：Join-476D058E-E953-4C2E-9456-188332E45400.rec

开始处理

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：

数据预处理：数据均衡
表：RealFrame_abalone.rec
数据均衡的目标列：Sex
数据均衡方法：Upsampling
结果命名：Equalization-9142F421-A93E-44FF-8F0F-33366C7DCE56.rec

根据该列的数据取值的分布情况进行均衡。

开始处理

4、值属性变换，针对数据类型不符合的情况：

数据预处理：值属性变换
选择表：RealFrame_abalone.rec

编辑列名称和类型

列名	数据类型	列名	数据类型	列名	数据类型
1 Sex	String	M	M	F	M
2 Length	String	0.72	0.73	0.775	0.505
3 DiaFeter	Numeric	0.575	0.555	0.57	0.39
4 Height	Numeric	0.215	0.18	0.22	0.115
5 Wholeweight	Numeric	2.1	1.6895	2.032	0.66
6 Shuckedweight	Numeric	0.8565	0.6555	0.735	0.3045
7 Visceraweight	Numeric	0.4825	0.1965	0.4755	0.1555
8 Shellweight	Numeric	0.602	0.4935	0.6585	0.175
9 Rings	Numeric	12.0	10.0	17.0	8.0

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

数据预处理：列过滤

*选择表： RealFrame_abalone.rec

*保留字段： 已选(7)

VisceraeMght	double
Shuckedweight	double
Wholeweight	double
Height	double
DiaFeter	double
Length	double
Sex	string

*未选(2)

Shellweight	double
Rings	double

*结果命名： ColumnFilter-B47F3B64-4CAB-4CEF-A66D-9D26E8E85D5D.rec

开始处理

6、增加序列号，对数据集某列排序，查看数据集：

数据预处理

数据预处理：增加序列号

*选择表： RealFrame_abalone.rec

*order by: Sex 升序

*结果命名： CreateIndex-E659A5AA-7F1E-48C2-937A-E25F58D9E8D9.rec

开始处理

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

数据预处理：联合(Union)

*左表： RealFrame_abalone.rec

*左表字段： 已选(3)

DiaFeter	double
Length	double
Sex	string

*右表： RealFrame_abalone.rec

*右表字段： 已选(3)

DiaFeter	double
Length	double
Sex	string

*未选(6)

Height	double
Wholeweight	double
Shuckedweight	double
Shellweight	double
Rings	double
VisceraeMght	double

*未选(6)

Height	double
Wholeweight	double
Shuckedweight	double
VisceraeMght	double
Shellweight	double
Rings	double

*去重:

*结果命名： Union-D142D5D-A9C9-4EE4-8E99-1FC178F93D0A.rec

开始处理

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