Li, Jun A

From: Lin, XiangX
Sent: Tuesday, January 26, 2016 2:58 PM
To: Li, Jun A
Cc: Zhang, Lijuan
Subject: TAP平台量化数据步骤

TAP 平台量化数据步骤

1 预处理
设置空值填补，填补为“32768”
2 上传 hdfs
3 连接服务器
4 创建 frame
5 数据提取
找出不属于选取范围的 feature
根据指定的 feature 列表进行删除操作；
6 量化分类
生成一个需要量化 feature 列表和一个不量化数据的列表
通过数组来指定列名处理
每处理一个 feature 新生产一个列，列名命名方式为：HCT -> Classify_HCT
不进行量化的类转换名称
7 填补替换
将整列数据提取到一个数组中
该数组去掉默认值“32768”，然后创建一个单独的 frame
根据创建的 frame， 执行中位数、均值、频次等操作；
根据计算值，对列进行再次转换，生成一个新列，列名命名方式为：Classify _HCT -> Fill_HCT
8 删除列
删除不是以 Fill_开头命名的列
9 列重命名
将 Fill_的列恢复：Fill_HCT -> HCT

Thanks and Regards

Sean Lin （Lin , Xiang）

cell: (+86) 135 85 198381, email: xiangx.lin@intel.com