**特征&标签工程文档**

本文档记录了对处理后的processed\_data进行充电过程划分（函数：split\_data）、标签工程（函数：cal\_score）和特征工程（函数：cal\_feature）。

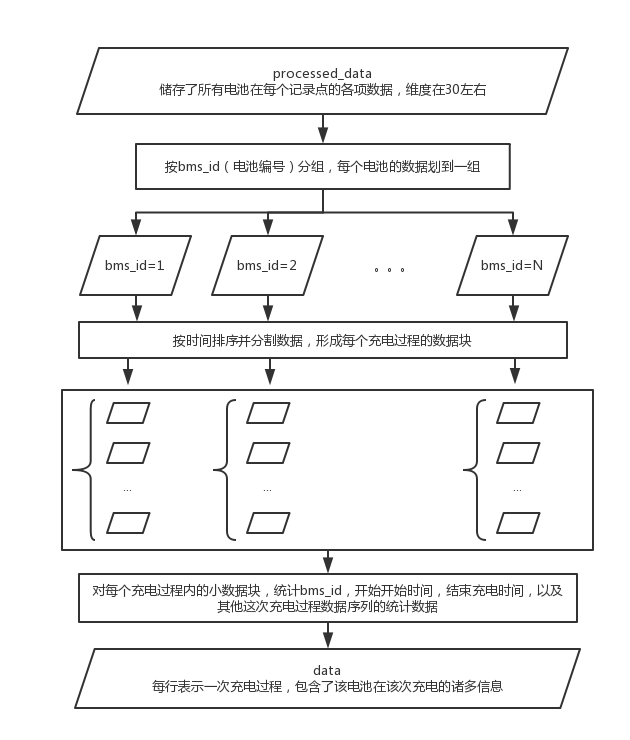
为了节省时间（split\_data是这个过程比较耗时的部分，因此对split\_data后的数据暂时保存为data.csv，方便修改cal\_score和cal\_feature ，来调整标签&特征工程的方法，便于优化模型）。

经过之前的数据清洗和数据处理步骤后，所得的数据被缓存为processed\_data。其中每一行代表一条充电信息记录，包括改电池的bms\_id，信息采集的时间点、此刻的电压电流、单体温度和电压和该电池的出场信息等，具体字段如下：



# 一、划分充电过程（split\_data）

本函数用于划分充电过程，主要的过程是：先按照bms\_id分组，得到每个电池的历史充电记录数据，然后按照time对改电池的历史充电数据进行排序；然后时间上从前到后遍历，发现前后的数据时间相差超过CHARGE\_TIMEGAP（单位为秒，现在代码里设置为300），就从这里进行分割，从而对每个电池的多次充电过程形成了对应的多个数据表；最后对每个小的数据表，统一计算一些统计量形成一行用于描述这次充电过程的多个数据，作为代表这一次充电过程的数据行，并入最终输出的data里。下图展示了这个流程：



具体的细节如下：

1. 过滤稀疏数据行：如果一行数据中的nan值（即空值）过多，则认为这一行数据信息量不大，对建模来说是噪声，因此直接舍弃。代码里设置DROPNA\_THRESH=12。
2. 数据格式整理：对数据格式进行整理，包括把时间列整理为时间戳，bms\_id和charger\_id整理为字符串。
3. 划分充电过程的最小时间间隔为CHARGE\_TIMEGAP，连续的充电记录时间差大于这个值时才进行划分，代码里暂时设置为CHARGE\_TIMEGAP=300秒。
4. 处理细分后的小数据块：经过检查，细分后的小数据块中很容易出现缺失值。因为一个小数据块是一个电池一段时间内的所有记录数据，因此每一行数据可以看做一个随时间变化的序列。对于一条序列上的缺失值，由于序列数据具有连续性，所以采取了前后向填充，即用前或后一个时间点的数据值填充。
5. 保留基础信息：对于每个小数据块，我们希望处理后的一行数据中包含对这个小数据块的信息。其中基础信息有：bms\_id，start\_time（充电开始时间），end\_time（充电结束时间），charger\_id和data\_num（充电记录条数）。
6. 计算容量表征信息：在电池领域，通常使用电量（单位：kwh）来表征某一时刻电池包含的能量，使用SOC（单位：%）表征此刻电池的容量状态，而某一时刻t的电池的容量可以用下式来描述：



我们将其记作‘score\_ca\_kwh’，我们对一个小数据块，计算其每一时刻的score\_ca\_kwh，并入数据列，用于后续计算。

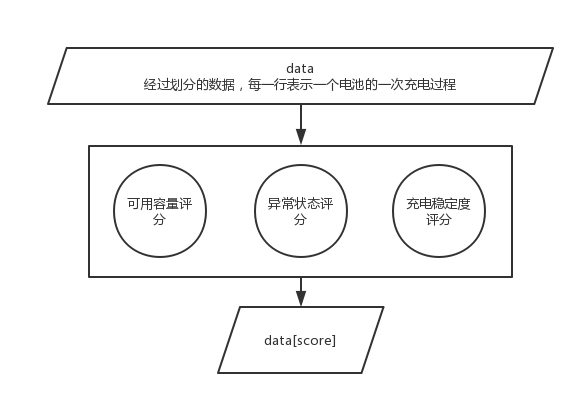
1. 计算差分、差分率和二阶差分：小数据块中的数据的每一行表示一条充电记录，每一列为电池的某个相关值的序列数据。对于电池的数据，其序列数据是随着时间变化的，比如电压和电流，为了能够更好地表达这种随时间变化的序列数据的信息，我们还需要每一列计算其一阶差分、一阶差分率和二阶差分等信息，才能够表达出这条序列数据随时间的变化情况。举个例子：电池的‘电压’随着充电的进行是在变化的，那么计算其一阶差分和一阶差分率可以描述它的变化情况，如果某一时刻电压发生了突变，则可以追踪电压的一阶差分是否过高，来判断电池是否有异常。
2. 序列数据转统计数据：小数据块中的数据的每一行表示一条充电记录，每一列为电池的某个相关值的序列数据，一般我们通过计算统计数据（均值，方差，最小值，最大值，中位数）的方式将序列数据转为一行的有限维度数据。举个例子：比如某一列记录了该电池在此次充电过程中的电压数据，该数据为一个长度为T的序列，我们对该序列计算有限个（设为M个）统计值（均值，方差，最小值，最大值，中位数），则该列数据转为M（M有限）个数值，那么对于一个有K列的小数据块，经过【序列数据转统计数据】的过程后，一个T\*K的小数据块将转为一个1\*（K\*M）的向量数据，从而实现向量化。代码中计算了均值，方差，最小值，最大值，中位数共5个统计数据。

# 二、标签工程（cal\_score）

现在业界对于电池的健康度并没有一个特别统一的定义，但是一般情况下，我们可以从“可用容量”、“异常状态”和“充电稳定度”。

其中，“可用容量”是指电池的最大可充电容量，电池在使用过程中，最大可充电容量会逐渐衰减，当最大可充电容量衰减到一定程度时，我们认为该电池不可再继续使用；“异常状态”是指电池充电过程中电压、电流、温度、单体电压等数值是否超出了限定值，如果超出，则认为是一次充电异常，这种异常情况频繁出现也表示电池的健康度不甚良好；“充电稳定度”是指电池充电的数据序列的稳定程度，举例说明：一个电池的电压在一次充电过程中出现了很激烈的波动，则其相比于电压不怎么波动的电池来说，我们认为它在这次充电过程时的健康度较低。

根据上面的分析，我们从这三个方面建立了对一个电池在一次充电过程中的健康度的打分机制，如下图所示：



具体细节如下：

1. 可用容量评分（占比50%）：使用data里的score\_ca\_kwh\_mean字段（表示的是之前计算的score\_ca\_kwh在一次充电过程序列数据中的均值），对所有数值计算0.01分位数和0.99分位数，小于0.01分位数的设为0.01分位数，大于0.99分位数的设为0.99分位数。然后对其归一化，作为“容量评分”=data[‘score\_ca’]。

2. 异常状态检测评分（占比10%）：电池的额定信息里包含了‘电压限值’、‘电流限值’，‘温度限值’和‘单体电压限值’，对data里的bp\_v\_max、bp\_i\_max、max\_st\_max和max\_sv\_max，检查其是否超出对应限值，超出则记0分，未超出记1分，用各项的平均值作为本项得分。

3. 充电稳定度评分（占比30%）：通过data里诸多描述数值波动情况的数据，做上下限的重新赋值，再做归一化，根据数值大小做评分，总的评分为所有项的平均值。用到的数据列有：

std\_sv/sv\_mean/max：描述了单体电压/温度的波动方差的均值和最大值。

bp\_v/i\_std/diff\_mean/diff\_max/diffrate\_mean/diffrate\_max：描述了电压和电流的波动方差，一阶差分均值和最大值，一阶差分率的均值和最大值。

cp\_p\_std/diff\_mean/diff\_max/diffrate\_mean/diffrate\_max：描述了功率的波动方差，一阶差分均值和最大值，一阶差分率的均值和最大值。

max/mean\_sv/st\_std/diff\_mean/diff\_max/diffrate\_mean/diffrate\_max：描述了最大单体电压/温度和单体电压/温度均值的波动方差，一阶差分均值和最大值，一阶差分率的均值和最大值。

4. 充电桩health字段（占比10%）：除了上面的评分机制，我现在还用了充电桩自己对健康度评分health字段，用其均值health\_mean来描述。

# 三、特征工程（cal\_feature）

原始数据中可用的和建模相关的数据列有：



上述数值在split\_dat中计算了对应的统计值，对其进行分析，对建模有贡献的如下：

* 1. 单体电池温度（st）

记录了每个时间点，不同电池（BMS编号）的不同电池箱（电池箱号）的不同分组（分组号）下多个（每个分组下大概100个单体电池，所以共有100个单体电池温度）。

我们定义一个充电次为一次时间差不超过1分钟的连续的数据记录（数据每2秒生成一次）。解释一下：就是电动车一次插到充电机开始充电后，数据中心就会开始记录数据，每2秒记录一次，所以当连续的两次记录时间差开1分钟以上时，我们就可以判定这是两次充电过程，期间必定经过了一次电源的拔出和重新插入。

我们首先以上面的规则对数据进行划分，对一个电池（BMS编号）得到了很多的充电次数据块。对于每个充电次，我们希望提取能够描述这次充电过程电池“健康度”的特征。

在一次充电次中，会有若干个检测时间点的数据（假设有T个），每个时间点有若干个单体电池的温度值（假设K个），我们对每个时间点，计算此时的K个单体电池温度值的若干统计值（比如最小值min，最大值max，均值mean，方差std，中位数median等，假设共M种统计值），则对每个时间点我们采集出了M个统计值来描述该时间点的单体电池温度，形成了T\*M的数据序列，我们按时间顺序从前到后排列。我们的目标是对这一整个充电次提取若干个特征。

对于序列数据，在机器学习建模时，常用的静态特征为：对原始值，一阶差分、二阶差分、一阶差分率、二阶差分率等计算统计量。对应到上述的T\*M序列数据，我们将使用对原始值、一阶差分、二阶差分、一阶差分率计算统计量作为特征。

特征名命名方式见下表：



举例说明：min\*ST\*diff\*mean= min\_ST\_diff\_mean，表示一次充电次下每个时间点的单体电池温度的最小值组成的长度为T的序列的一阶差分值的均值。

* 1. 单体电池电压（sv）

处理方式类似于“单体电池电压记录”，特征名命名方式见下表：



举例说明：std\*SV\*diffrate\*min= std\_SV\_diffrate\_min，表示一次充电次下每个时间点的单体电压温度的方差值组成的长度为T的序列的一阶差分率值的最小值。

* 1. 电池充电过程

有用的数据基本只有‘SOC’、‘剩余时间’、‘电压’和‘电流’，后面的‘单体最低/高温度/电压’数值不准，而且上面的表更详细处理了，所以不做处理。

和“电梯电池温度”的处理方式差不多，也是先按时间差开1分钟划分成充电次，先对每次内很多数据点统计，再对很多个充电次分别做特征。

特征名命名方式见下表：



举例说明：median\*V\*diff2\*mean= median\_V\_diff2\_mean，表示一次充电次下每个时间点的电压值的中位数组成的长度为T的序列的二阶差分值的均值。

对于提取出的特征，还可以选择做特征白化：

1. 归一化

由于不同数据之间的量纲不同，而机器学习模型通常是对量纲比较敏感，所以在建模之前，需要对已提取的特征做归一化。

对特征向量 （假设共L个特征），令 ，即为白化后的特征向量。

2. 标准化

不同数据间除了量纲不同，通常还具有不同的均值和方差，所以需要将特征向量处理为均值为0，方差为1的标准化数据，使得噪声数据对模型的影响尽可能小。

对特征向量 （假设共L个特征），令 ，即为标准化后的特征向量。