**前言**

毕设分为2个部分：负荷识别和用户行为挖掘，在UKDALE数据集上开展（REDD数据太少）。以30分钟为窗口，使用模型识别设备开启事件，然后在识别结果上进行挖掘。

**负荷识别**

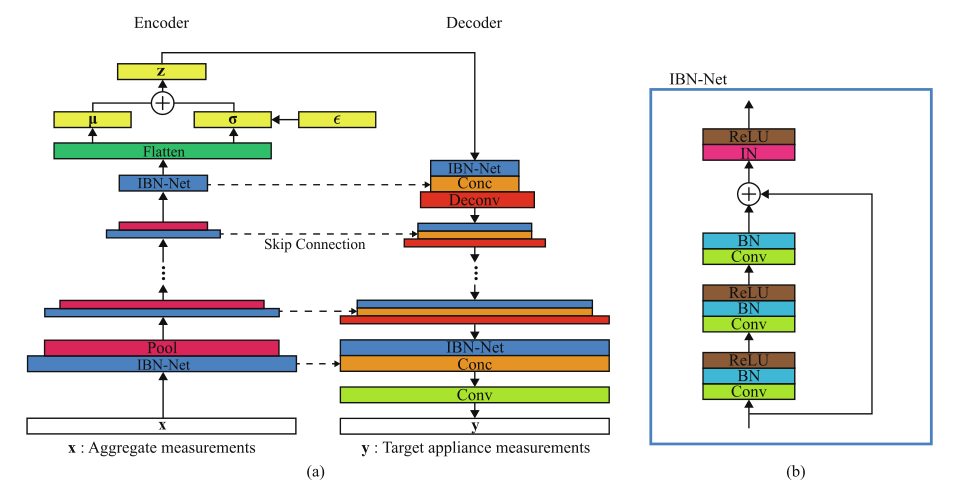
负荷识别的目标效果是：为每个设备训练一个模型，输入一段总线功率（下采样之后300个点：30分钟），通过模型得到该期间内设备是否存在启动事件（本质是一个二分类问题）。

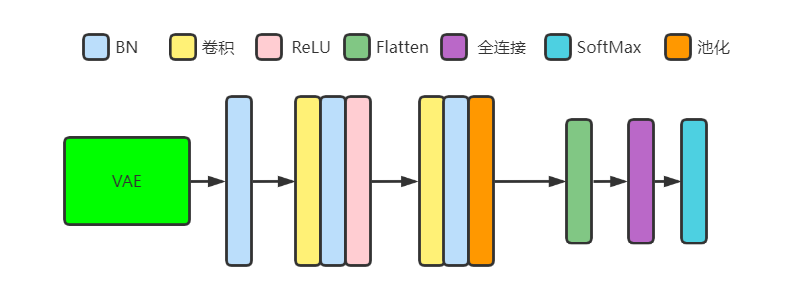
**数据处理**

1. 下采样、缺失值填充、输入归一化
2. 数据过滤：对于一个窗口，如果期间内设备没有启动事件（label == 0），则以0.9的概率删除该窗口

**模型结构**

以VAE为基础，在其结果上进行识别。





**训练**

1. 学习率：[10-3,10-4,10-5]，共20轮
2. 将输入数据扩展至320个点，也就是两边各扩展10个点，没有的话以0进行填充
3. 损失：KL散度+分解均方误差+识别交叉熵误差

**结果**

结果在很大程度上取决于分解的效果：

有效电器：kettle、washing\_machine、dishwasher

无效电器：microwave、小功率电器

**用户行为挖掘**

* 关联规则
* 设备与时间的联系
* 设备使用频次

**关联规则挖掘**

**负荷识别**

每天根据模型识别结果得到以下表格：

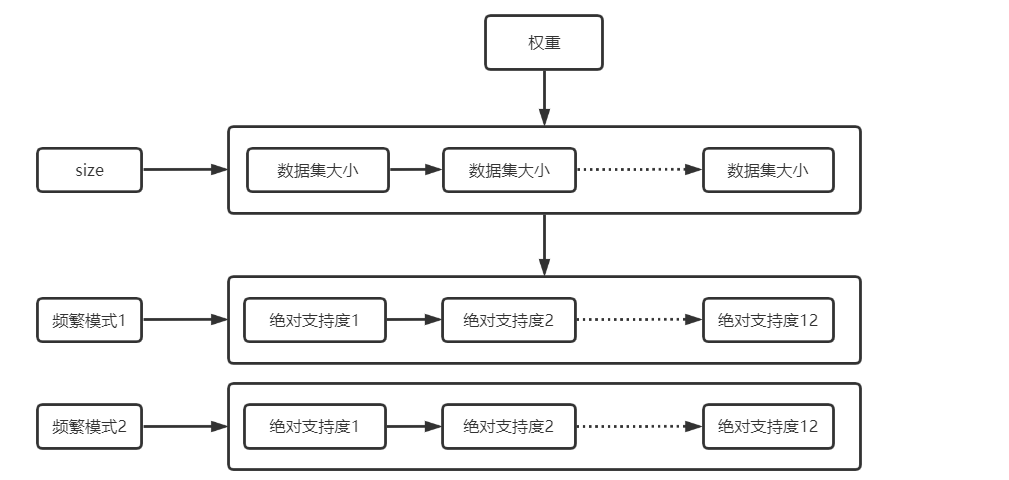
|  |  |  |  |
| --- | --- | --- | --- |
| 设备  时间 | amp | kettle | tv |
| 0 | 0 | 1 | 0 |
| 1 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 |
| … | … | … | … |
| 46 | 0 | 0 | 1 |
| 47 | 1 | 0 | 1 |

**频繁模式数据库**

频繁模式数据库记录所有项集每个月的绝对支持度，总共记录1年的数据。

* 所有项集：不包括没出现过的项集，即要求绝对支持度大于0
* 绝对支持度：就是对应项集出现的次数

每天根据识别结果对数据库进行一次更新。



**关联规则生成**

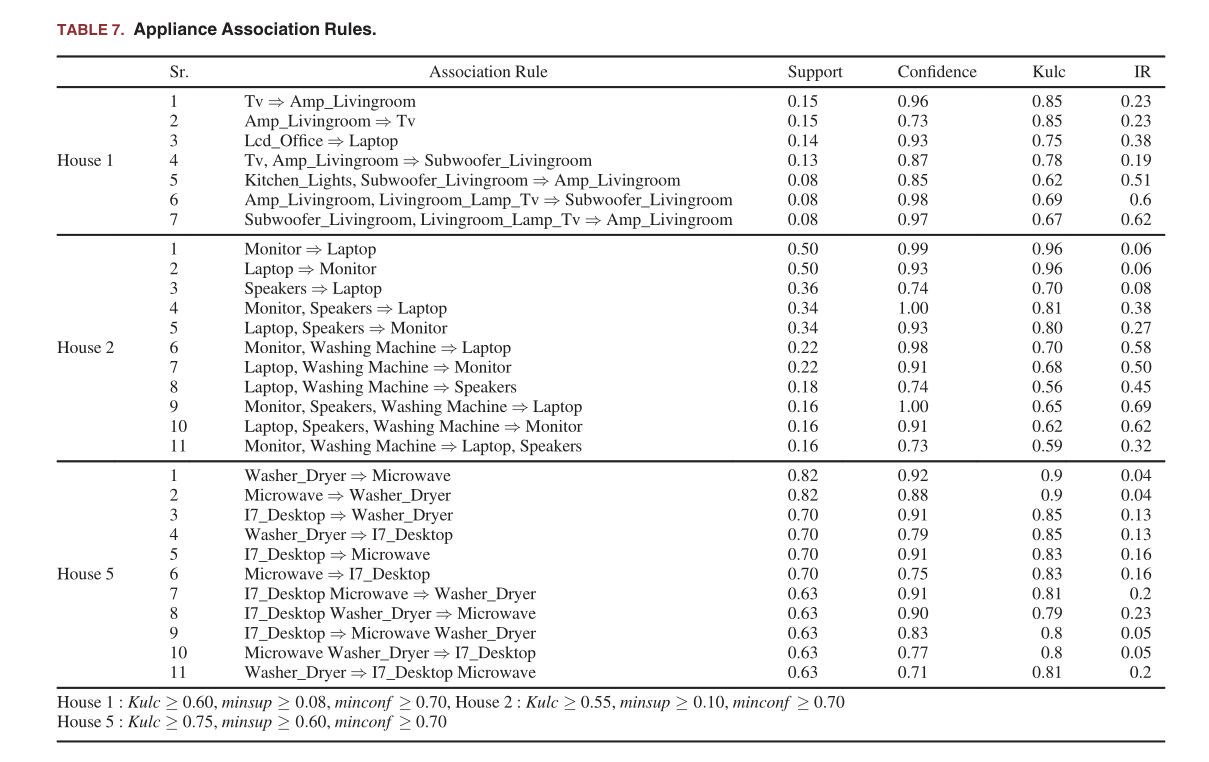
1. 根据最小支持度，过滤掉一部分频繁项集（即认为其是非频繁的），最小支持度设置在0.1到0.01是比较合适的。
2. 然后，在频繁项集的基础上生成关联规则：这部分与Apriori算法没有什么区别，就是一个排列组合，然后计算置信度
3. 最后，根据设置的最小置信度过滤关联规则，最小置信度设置为0.8到0.9比较合适。

**结果**

这部分数据直接来自标签，针对UKDALE house1的amp\_livingroom、kettle、tv进行挖掘，结果如下：



别的论文已经进行了挖掘，其结果如下：



**问题**

1. 现阶段问题

* 最主要的问题就是模型对很多电器是无效的（单看准确率和mae还是不错的），而那些有效的电器之间并没有关联关系
* 划分窗口的方式可能会截断事件，造成误识别
* 设备的功率曲线看起来不是很工整，有时会突然出现一个尖峰或者深谷，应当将其处理掉，但是最小开启/关闭时间不好确定

1. 其他问题

关联规则挖掘很浅，只能确定有关联；以30分钟为窗口可能会截断频繁模式，即设备位于两个窗口，但其实使用间隔小于30分钟，那么，它们应该是有关联的。针对此问题，需要模型输出事件在窗口内的具体时间，然后再开展进一步工作。

备注：

{1，2，7，8，2，3，7，8，8} => {1，2，2，3}、{7，8，7，8，8}

关联规则

冬天：TV--->暖手器

夏天：TV-x->暖手器

对比，确定识别和分解哪个是因。