论文封面

论文题目 家用电器用户行为分析及事件识别

题号 （填写所选题号A或B）

作者

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 班级 | 学号 | 小组 |
| 梁衡 | 云计算16-2 | 201611061055 | 8组 |
| 胡军 | 云计算16-2 | 201611061069 | 8租 |
| 赵瑞麒 | 云计算16-2 | 201611061067 | 8组 |

家用电器用户行为分析及事件识别

# 摘要

家用企业若能深入了解不同用户群的使用习惯，开发新功能，就能开拓新市场。厂商可从热水器智能操作和节能运行等多方面对产品进行优化。

热水器厂商根据洗浴事件识别模型，对不同地区的用户的用水进行识别，根据识别结果比较不同客户群的客户使用习惯、加深对客户的理解等。从而，厂商可以给不同的客户群提供最适合的个性化产品、改进新产品的智能化的研发和制定相应的营销策略。

对于问题1，由于不同地区的人们用热水器的习惯不同，以及不同季节的时候使用热水器时停顿的时长也可能不同，我们建立一个用水事件阈值寻优模型来设置一个合理的阈值T，用这个阈值区分同一事件中的停顿和不同用水事件之间的间隔，水流量为0的状态记录之间的时间间隔如果超过阈值T，从该段水流量不为0的状态记录向前找到最后一条水流量不为0的用水记录作为上一次用水事件的结束，向后找到水流量不为0的状态记录作为下一个用水事件的开始，因此，一次完整的用水事件划分完毕。

对于问题2，通过三个条件：一次用水事件中总用水量（纯热水）小于y 升、水时长小于100秒、用水时长小于120秒筛选出非常短暂用水事件，剩余的洗浴事件称为“候选洗浴事件”。再构造4个指标：时长类指标、频率类指标、用水量化指标、用水波动指标，利用这四个指标中的用水的用途、用水开始时间、用水结束时间等属性的用水日志来构建专家样本，通过BP神经网络算法建立一个洗浴识别模型，再调用此模型对专家样本数据进行模型检验，得出识别准确率为90.5%，所以利用此模型对题目所给数据进行预测，并将结果保存。

**关键字：**阈值寻优；BP神经网络模型；数据挖掘；洗浴识别模型；

# 一、问题重述

随着人们生活质量的提高，热水器走进了老百姓的家门，不但满足了人们基本的洗浴热水需求，还解决了洗手、做饭、刷牙洗菜等基本生活用水问题。

基于热水器采集的时间序列数据，将顺序排列的离散的用水时间节点划分为不同大小的时间区间，每个区间视为一次完整的用水事件，并以热水器一次完整用水事件作为一个基本事件，再把热水器采集的时间序列分为独立的用水事件，进一步识别出其中的洗浴事件。

问题包含以下两步：

1.根据热水器采集到的数据，划分一次完整用水事件；

2.在划分好的一次完整用水事件中，识别出洗浴事件。

# 二、问题分析

## 2.1 问题1的分析

问题1要求如何区分/定义一次完整的用水事件，在用水状态记录中，水流量不为0表明用户正在使用热水，水流量为0时用户用热水发生停顿或者用热水结束，一次用水事件之间的停顿通常小于两个不同用水事件之间的间隔，所以我们需要设置合理的阈值，以区分同一用水事件中的停顿和不同用水事件之间的间隔，根据这个间隔可以计算出不同的用水事件划分

由于不同地区的人们用热水器的习惯不同，以及不同季节的时候使用热水器时停顿的时长也可能不同，固定的停顿时长阈值对于某些特殊的情况的处理是不理想的，存在把一个事件划分为两个事件或者把两个事件合为一个事件的情况，所以我们需要建立一个阈值寻优模型以解决因时间变化和地域不同导致阈值存在差异的问题。

## 2.2 问题2的分析

问题2要求我们如何识别洗浴事件，由于洗浴事件的识别是建立在一次用水事件识别的基础上，识别出哪些是洗浴事件，所以我们需要先筛选出非常短暂的用水事件，剩余的用水事件即为“候选洗浴事件”，我们还需要构造时长类指标、频率类指标、用水量化指标、用水波动指标及其若干属性，根据BP神经网络建立洗浴事件识别模型，接着调用洗浴时间识别模型，对实时监控的热水器流水数据进行洗浴时间自动识别。

# 三、模型假设

1.假设检测热水器不存在漏水漏电问题；

2.假设采用同种热水器进行采样收集数据；

3.假设水温测量无偏差，期间热水器不散热；

# 四、符号说明

|  |  |
| --- | --- |
| 名称 | 符号 |
| 状态记录i |  |
| 时间间隔阈值 | T |
|  |  |
| 与*i*点的斜率的绝对值  *i∈* {*B,C,D,E*} |  |
| 五个点的斜率之和的平均值 | K |
| 任意两点(*x1,y1*) (*x2,y2*)， 的斜率的绝对值 | *k* |
| *i*点的坐标  *i∈* {*A,B,C,D,E*} | *(xi,yi)* |

# 五、模型建立及求解

## 5.1.数据的预处理

### 5.1.1数据规约

分别对数据的属性和数值进行规约，因“热水器编号”、“有无水流” 、“节能模式”对建模无作用，可以去除；当热水器“开关机状态”为“关”且水流量为0时，说明热水器不处于工作状态，数据记录可以规约掉。

归约后的部分数据如下：

表1 热水器各时间点的属性及数据

| 发生时间 | 开关状态 | 加热中 | 保温中 | 实际温度 | 热水量 | 水流量 | 加热剩余时间 | 当前设置 温度 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|
| 20141019161042 | 开 | 开 | 关 | 48°C | 25% | 0 | 1分钟 | 50°C |
|
| 20141019161106 | 开 | 开 | 关 | 49°C | 25% | 0 | 1分钟 | 50°C |
| 20141019161147 | 开 | 开 | 关 | 49°C | 25% | 0 | 0分钟 | 50°C |
| 20141019161149 | 开 | 关 | 开 | 50°C | 100% | 0 | 0分钟 | 50°C |
| 20141019172319 | 开 | 关 | 开 | 50°C | 50% | 0 | 0分钟 | 50°C |
| 20141019172321 | 关 | 关 | 关 | 50°C | 50% | 62 | 0分钟 | 50°C |
| 20141019172323 | 关 | 关 | 关 | 50°C | 50% | 63 | 0分钟 | 50°C |

### 5.1.2 缺失值处理

首先本案例中存在用水数据状态记录缺失的情况，划分一次完整用水事件时，需要一个开始用水的状态记录和结束用水的状态记录。在划分一次完整用水事件时，发现数据中存在没有结束用水的状态记录情况：热水器状态发生改变，第5条状态记录和第7条状态记录的时间间隔应该为2秒，而表中两条记录间隔为1小时27分28秒。因存在网络故障等原因，导致用水数据状态记录缺失的情况，需要对缺失的数据状态记录进行添加。其添加方法：用水状态记录缺失的情况下，填充一条状态记录使水流量为0，发生时间加2秒，其余属性状态不变。

## 5.2探索分析

用水停顿时间间隔定义为一条水流量不为0的流水记录同下一条水流量不为0的流水记录之间的时间间隔。

根据现场实验统计，两次用水过程的用水停顿的间隔时长一般不大于4分钟。

为了探究用户真实用水停顿时间间隔的分布情况，统计用水停顿的时间间隔并作频率分布直方图。

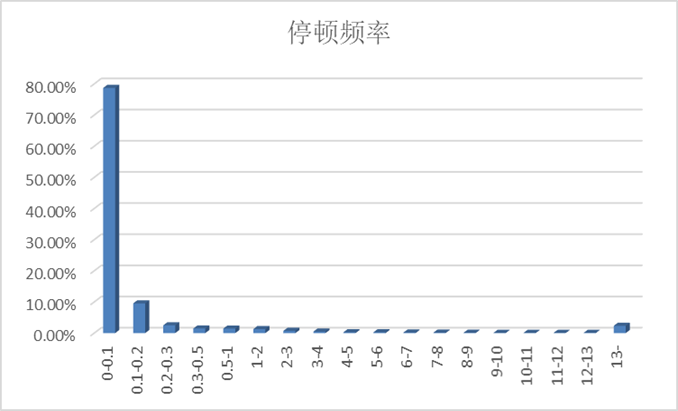


图1 用水停顿时间间隔表

停顿时间间隔为0~0.3分钟的频率很高，根据日常用水经验可以判断其为一次用水时间中的停顿

停顿时间间隔为6~13分钟的频率较低，分析其为两次用水事件之间的停顿间隔。

两次用水事件的停顿时间间隔分布在3~7分钟与现场实验统计用水停顿的时间间隔近似。

## 5.3 挖掘建模

### 5.3.1 建模目标

对热水器用户的洗浴事件进行识别——>需要从原始数据中识别出哪些状态记录是一个完整的用水事件（包括洗脸、洗手、刷牙、洗头、洗菜、洗浴等）——>再识别出用水事件中的洗浴事件

### 5.3.2建模步骤

1.一次完整用水事件的划分

2.用水事件阈值寻优

3.候选洗浴事件

4.属性构造

5.神经网络训练

6.识别洗浴事件

### 5.3.3建模过程的实现

**a.一次完整用水事件划分**

用水记录状态中：水流量不为0表明用户正在使用热水，水流量为0时用户用热水发生停顿或者用热水结束，一次完整用水事件中可能会发生多次停顿。所以我们通过阈值来区分同一事件中的停顿和不同用水事件之间的间隔。

设置阈值步骤如下：

1）读取数据记录，识别到第一条水流量不为0的数据记录记为R1，按顺序识别接下来的一条水流量不为0数据记录为R2 。

2）若*gap*i>T， Ri+1与Ri及之间的数据记录不能划分到同一次用水事件，将 Ri+1记录作为读取数据记录的开始，返回步骤1）；

若 *gap*i<T，则将Ri+1 与 Ri之间数据记录划分到同一次用水事件, 并记录接下来的水流量不为0状态记录为 Ri+2。

3）循环执行步骤2）直到数据记录读取完毕，结束事件划分。

**b. 用水事件阈值寻优模型**

考虑到不同地区的人们用热水器的习惯不同，以及不同季节的时候使用热水器时停顿的时长也可能不同，固定的停顿时长阈值对于某些特殊的情况的处理是不理想的，存在把一个事件划分为两个事件或者把两个事件合为一个事件的情况。

本案例建立了阈值寻优模型来更新寻找最优的阈值，这样可以解决因时间变化和地域不同导致阈值存在差异的问题。我们利用程序算出各个阈值所划分的事件个数如下表所示：

表2 阈值与事件个数

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 阈值（min） | 2.25 | 2.5 | 2.75 | 3 | 3.25 | 3.5 | 3.75 | 4 | 4.25 | 4.5 | 4.75 | 5 |
| 事件个数 | 650 | 644 | 626 | 602 | 588 | 565 | 533 | 530 | 530 | 530 | 522 | 520 |
| 阈值（min） | 5.25 | 5.5 | 5.75 | 6 | 6.25 | 6.5 | 6.75 | 7 | 7.25 | 7.5 | 7.75 | 8 |
| 事件个数 | 510 | 506 | 503 | 500 | 480 | 472 | 466 | 462 | 460 | 460 | 460 | 460 |

阈值与划分事件个数的散点图如下图所示：



图2 阈值与划分事件个数散点图

图中某段阈值范围内，下降趋势明显，说明在该段阈值范围内，用户的停顿习惯比较集中。

如果趋势比较平缓，则说明用户的停顿热水的习惯趋于稳定，所以取该段时间开始的作为阈值，既不会将短的用水事件合并，又不会将长的用水事件拆开。

曲线在图2中方框趋于稳定时，其方框开始的点的斜率趋于一个较小的值。我们作出了这几个点的斜率图如下图所示：



图3 用水停顿斜率图

根据图3，我们可以构建一个阈值优化的方法：

当存一个阈值的斜率指标K<1 时，则取阈值最小的点A （可能存在多个阈值的斜率指标小于1）的横坐标 *xA*作为用水事件划分的阈值。

K<1中的1是经过实际数据验证的一个专家阈值。

当不存在K<1 时，则找所有阈值中斜率指标最小的阈值；如果该阈值的斜率指标小于5，则取该阈值作为用水事件划分的阈值；如果该阈值的斜率指标不小于5，则阈值取默认值的阈值4分钟。

斜率指标小于5中的5是经过实际数据验证的一个专家阈值。

**c. 筛选得“候选洗浴事件”**

洗浴事件的识别是建立在一次用水事件识别的基础上，也就是从已经划分好的一次用水事件中识别出哪些一次用水事件是洗浴事件。

首先用3个比较宽松的条件筛选掉那些非常短暂的用水事件，剩余的洗浴事件称为“候选洗浴事件”。

三个条件是“或”的关系，如下：

一次用水事件中总用水量（纯热水）小于y 升.

用水时长小于100秒。

总用水时长小于120秒。

提示：经过实验分析，热水器设定温度为50摄氏度时，一次普通的洗浴时长为15分钟，总用水时长10分钟左右，热水的使用量为10~15升。

为不影响特殊的短暂的洗浴事件，以及考虑到夏天用的热水较少，放宽范围。

假定热水器在设定温度为50摄氏度时，一次洗浴的总热水使用量为5升，同时取洗浴温度的均值为39摄氏度。来计算热水器不同设定温度下的热水使用量阈值。

|  |  |  |  |
| --- | --- | --- | --- |
| 洗浴用水温度 | T(39摄氏度) | 设置温度 | X（摄氏度） |
| 自来水水温 | C(摄氏度) | 设定温度为时的用水量 | Y（升） |
| 自来水注水量 | M（升） | 50摄氏度时的用水量 | V（5升） |

假定每次洗浴习惯变化不大且热水器热水水温恒定，则每次洗浴使用的热水的热量应该趋近于一个定值。如果热水器设定温度 X调高使热水器水温变高，则一次洗浴使用的热水量就减少；相反，则使用的热水量就增多。

假设两次洗浴事件热水和冷水混合后的花洒出水温度恒为T 摄氏度，总用水量不变且为 M+V升，根据热量守恒建立方程组（13-2）。



（1）式是50摄氏度的热水V升与M 升C 摄氏度自来水混合得到M+V 升 T摄氏度的洗浴用水的热守恒公式。

（2）式是 X摄氏度的热水Y 升与M+V-Y 升C 摄氏度自来水混合得到M+V 升T 摄氏度的洗浴用水的热守恒公式。

从而得出 Y与X 、C 、V 之间的关系：



**d.属性构造**

1、时长类指标



由上表可知，在20:00:10时热水器记录到的数据还没有用水，而在20:00:12时热水器记录的有用水行为。

用水开始时间在20:00:10~20:00:12之间，考虑到网络不稳定导致网络数据

传输延时数分钟或数小时之久等因素，取平均值会导致很大的偏差，综合分析构建“用水开始时间”为起始数据的时间减去“发送阈值”的一半，发送阈值是指热水器传输数据的频率的大小

同理构造用水结束时间、停顿开始时间、停顿结束时间等

“用水时长A”是“用水开始时间”到“停顿开始时间”的间隔时长

一次用水事件中“用水时长”为各段用水时长之和

同理构造总用水时长、停顿时长等。

2、频率类指标

统计一次用水事件中各种用水操作的频率。

3、用水量化指标

总用水量定义为：在水流量不为0时，一次用水事件如表 13‑10中每条状态记录的水流量与下一条状态记录的时间间隔的乘积；平均水流量定义为总用水量与用水时长的商。

4、用水波动指标

“水流量波动”指标定义为当前水流的值与平均水流量差的平方乘以持续时间的总和除以总的有水流量的时间。同理构造温度波动、热水量波动、停顿时长波动等指标。

# 六、模型评价

## 6.1 模型优点

本文根据建模样本数据和用户记录的用水日志，通过阈值寻优，较合理的划分了用水事件，并最终建立了BP神经网络模型识别洗浴事件。该模型的识别率较高，并且具有一定的容错能力,特别适合于求解内部机制复杂的问题。

## 6.2 模型缺点

由于BP神经网络自身的局限性，导致收敛速度慢，容易陷入局部极值。针对此问题，可用RBF神经网络对模型进行改进。

### 6.3 模型推广

1)问题中对于热水器使用状况的分析采用了阈值模型，这种模型清楚地划分出了一次完整用水事件。所以这类模型可以应用推广到划分一类事件的问题中，用以区分一类事件。

2)问题中用到的阈值优化寻值模型对于优化问题的处理有一定的帮助，这种思想对其他数学问题及一般的模型仍能使用。

# 七、参考文献

[1] 司守奎 孙兆亮,数学建模算法与应用(第2版）,北京:国防工业出版社,2017.

[2] 吴建国，数学建模案例精编，北京，中国水利水电出版社，2005.

[3] 汪莹, 基于数据挖掘的安全管理信息系统研究,中国矿业大学学报,2014年02期

[4] evillist,数据挖掘的一般过程

https://blog.csdn.net/evillist/article/details/73275188 2018.7.27

[5] sunxinyu , RBF神经网络与BP神经网络优缺点比较

<https://blog.csdn.net/sunxinyu/article/details/76598446>, 2018.7.27