序号： 2020121001

安顿·血压二期项目上线报告

报告内容： 安顿·血压项目（二期）

固定电话：

移动电话：

电子邮箱：

北京雪扬科技·技术中心算法部

二〇二〇年十二月制

[一、 目的 3](#_Toc4575)

[二、 适用范围和声明 3](#_Toc9308)

[三、 数据采集流程 2](#_Toc2311)

[1. 自动采集模式（使用康康血压计） 2](#_Toc23765)

[2. 手动采集模式 3](#_Toc6506)

[四、 数据获取方法 5](#_Toc17248)

[1. 用户测量的真实血压数据 5](#_Toc20301)

[2. 血压数据对应的PPG特征数据 7](#_Toc21320)

[五、 数据与特征提取 8](#_Toc16516)

[1. 数据说明 8](#_Toc31740)

[2. 特征提取 14](#_Toc13311)

[六、 算法与策略 16](#_Toc24281)

[1. 模型结构 16](#_Toc27834)

[2. 损失函数 17](#_Toc10953)

[3. 正则化 17](#_Toc27827)

[4. 模型融合 17](#_Toc22817)

[5. 出值策略 18](#_Toc18949)

[6. 线上部署方式 19](#_Toc11967)

[7. 模型Retrain 19](#_Toc3212)

[七、 算法优化 19](#_Toc30883)

[1. 基于算法模型的优化策略 19](#_Toc26521)

[八、 当前算法存在的问题 21](#_Toc30205)

[九、 项目未来的发展方向 21](#_Toc6701)

[1、 工程实现 21](#_Toc23543)

[2、 硬件角度 22](#_Toc30263)

## 目的

本文档主要提供安顿技术中心算法部门阶段性研究方向和内容报告，其中详尽提供相关研究内容、具体算法、实际测试结果、应用场景和面临挑战等信息。

本期上线的是针对于个人用户的血压模型,是对于第一期血压模型的补充。第一期上线的血压模型是通用血压模型，目的是覆盖大部分线上用户，但是由于数据采集的局限性，第一期的血压模型在少数用户身上，测量血压时存在较大误差，如果按照第一期血压模型采集数据、训练模型方法，使用此用户采集的少量数据进行模型的重新训练时，模型并不一定能学习到此用户的血压特征，而且模型的重新训练还会对之前其他用户的血压预测产生影响。综合考虑，针对于少量用户的血压预测偏差较大的情况，使用此用户采集的少量血压数据,训练一个独立的个人模型,此模型专门用来对个人用户的血压进行预测。

## 适用范围和声明

适用于北京雪扬科技有限公司技术中心算法部门内部流通，其余各部门和个人不得以任何形式外传（包括原件、部分粘贴、拍照截图和口述内容等），仅以存档供查阅和作为总结、指导使用，如阅读此文档当遵守适用范围。

## 数据采集流程

个人用户的血压数据采集比较灵活，可以使用康康血压计，也可以使用用户自己的其他型号的血压计进行测量。如下的采集流程1和2

### 自动采集模式（使用康康血压计）

#### 自动采集模式

用户只需要佩戴康康血压计，把血压计调到自动采集模式，康康血压计会在固定的时间周期采集血压数据（白天是半小时采集一次，晚上是20分钟采集一次），当康康血压计开始充气加压采集时，用户需要保持静止状态，最好是保持坐姿，手臂与心脏保持水平。

#### **采集对象**

反馈血压数据失准的用户。

#### **采集要求**

佩戴人尽可能全天佩戴，佩戴期间至少有1晚的夜间数据。用户佩戴康康血压计期间需要同时佩戴安顿手表，而且血压计和手表要佩戴在一只手臂上。  
自动模式采集期间,此血压计不能用于其他人的血压采集。

#### **准备工作**

(1)记录血压计的编号和领取人的关系；

(2)帮助领取血压计的人员绑定血压计；

(3)统计佩戴人的详细信息，用于下发采集的使用；

(4)与康康协调导出血压计数据；

(5)通过血压计与微信关联的Family Id来获取血压计数据；

#### **血压计的回收**

1. 在归还血压计时确定是否有新领取护心手表，如有领取需一并归还；
2. 归还血压计后，血压计是否为绑定状态；
3. 血压计包装是否齐全；

#### **采集数据**

在用户需要提供使用的康康血压计的ID号以及自动模式采集的时间段(开始时间,结束时间),我们自动从后台获取血压计测量的数据。

### **手动采集模式**

用户可以使用康康血压计(调整到单次采集模式)，或者用户自己家里的血压计(水银血压计，电子血压计)，配合安顿手表的手动测量模式进行血压的采集。

#### 2.1采集对象

反馈血压数据失准的用户。

#### 2.2采集要求

1. 佩戴好安顿手表；
2. 采用标准坐姿，在佩戴安顿手表的手臂上进行血压的测量，可以使用康康血压计，或者其他经过医疗认证的血压计；
3. 测量完成一次需要记录测量的时间（精确到分），及测量的高压，低压值；
4. 每当使用血压计测量完成一次血压，需要使用安顿手表进行一次手动血压测量（目的是为了采集此时用户的生理数据），测量间隔时间不超过8分钟；
5. 一次完整的数据采集包含一次血压计采集，一次安顿手表采集，用户需要分为早、中、晚，每个时间采集3组完整的数据采集，每次完整的数据采集时间间隔在10分钟以上；
6. 如果某个时间没有采集数据，也可以在一天内的其他时间进行采集，如果是高血压患者，并且有服用降压药的，应该在服药前、服药后半小时分别进行测量；
7. 保证一天内采集10组左右的完整数据；

#### 2.3血压计申请（康康血压计）

* + 合伙人可以通过填写表单的方式进行申请；
  + 用户反馈自己血压数据失准，会根据用户情况，联系当地合伙人申请血压计给用户进行采集；
  + 邮寄血压计时查看血压计是否为绑定状态；

#### 2.5数据审核

* + 用户测量完成后，把数据通过电话的发给安顿客服，客服会按照对应的时间把血压数据录入安顿系统；
  + 安顿系统会自动判断某条血压数据是否有效（系统查找血压对应时间点的前后8分钟内是否有手表测量的PPG特征数据,如果有则认为此条血压数据有效,否则认为无效）；

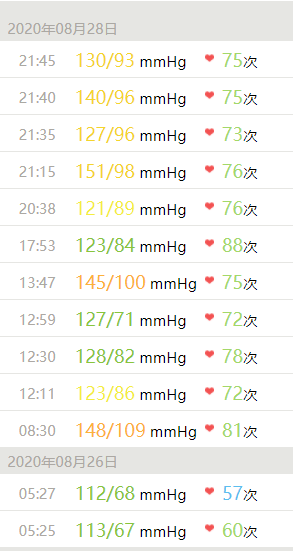
1. 数据获取方法

血压数据采集脚本共分为2个模块，用户测量的真实血压数据、用户测量血压当天,手表测量的PPG特征数据。通过查找每条血压数据测量时,前后8分钟的PPG特征数据,组成训练数据。

### 用户测量的真实血压数据

* 1. 对于用户使用康康血压计采集的血压数据,我们可以通过血压计的ID号,以爬虫的方式爬取此用户采集的数据, 需要用户提供使用的康康血压计设备的ID号, 以及测量血压的时间段(开始时间,结束时间);

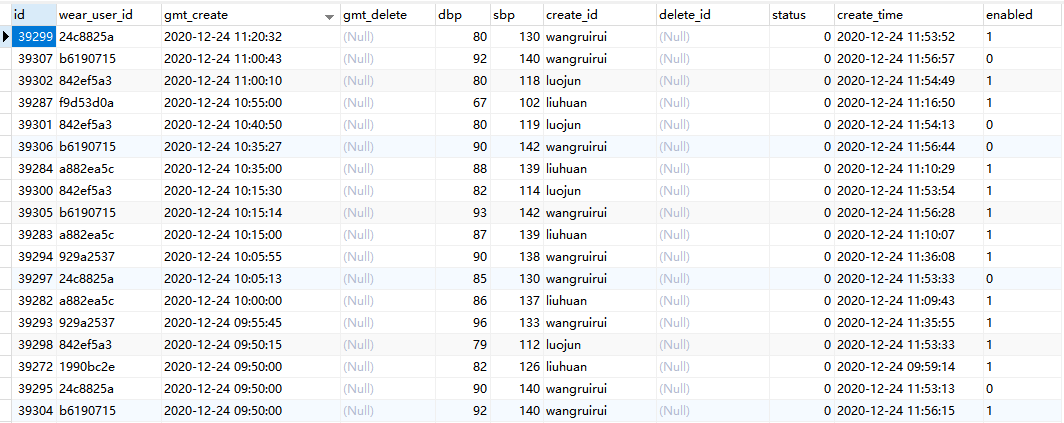
康康官网的血压数据如下:



* 1. 如果是用户使用其他型号的血压计测量的数据,通过电话的方式报送给安顿客服工作人员,工作人员会录入安顿系统,我们只需要从后台数据库读取此用户的血压数据即可,

安顿后台记录的用户测量的血压数据：



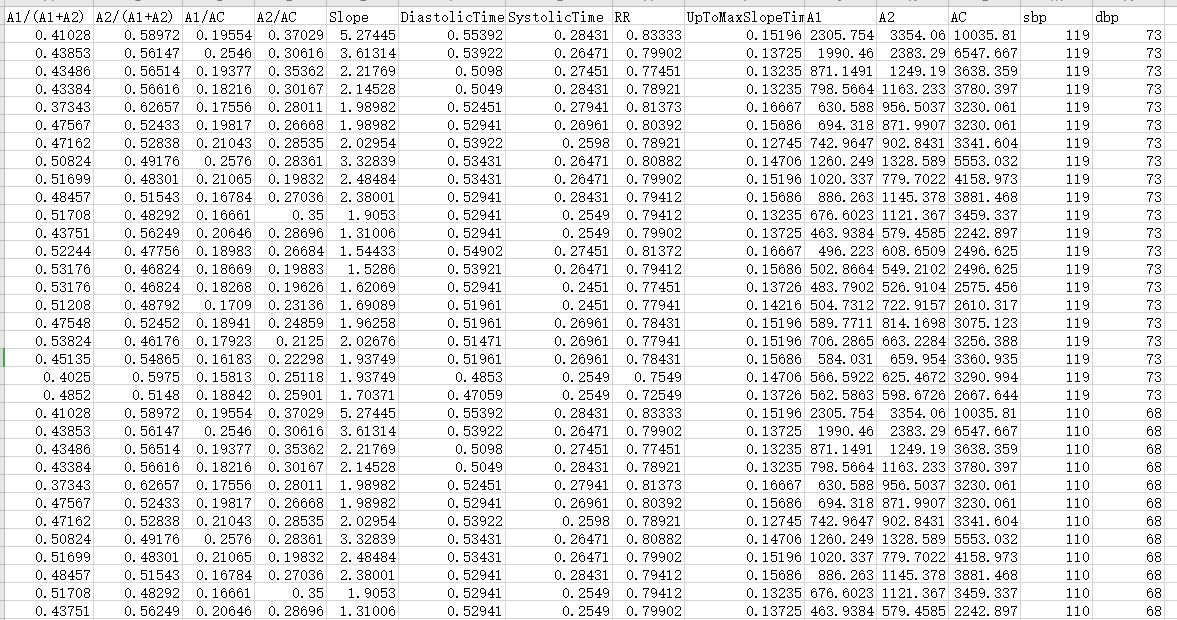
数据库保存的用户血压数据如下:  


直接通过wear\_user\_id即可获取某个指定用户测量的所有血压数据。

### 血压数据对应的PPG特征数据

2.1、首先查询用户血压数据当天的所有PPG特征数据，此数据从后台数据库通过wear\_user\_id和date(日期)即可查询;

2.2、把每条用户测量的血压数据，找出其对应的前后8分钟内的所有PPG特征数据（此处假设前后8分钟内用户的血压是平稳的），整理好之后，所有的PPG特征数据作为feature特征值,血压值作为训练目标label,整理后的数据如下图:



## 数据与特征提取

### 1. 数据说明

#### 1.1 原始数据说明

一个典型的PPG信号周期, 一个信号周期包括了上升支和下降支两个部分, 上升支对应收缩期, 心室快速射血, 动脉中的血液量迅速增加; 下降支对应舒张期, 心室射血量下降, 血液经过毛细血管网络流入静脉, 动脉中的血液量减少. PPG波形的下降支又出现一个波峰, 此波峰称为重博波. 重博波的产生是由于心脏舒张期, 大动脉根部由于心室射血量减少而造成血压降低, 进而出现血液回流而造成的. 在重博波之后动脉瓣膜快速关闭, 血压流向恢复正常, PPG信号沿下降支继续下降.

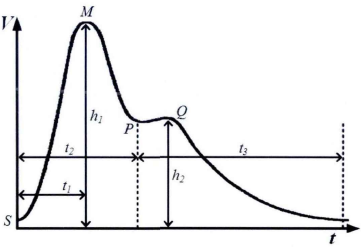


图. PPG波形及其特征参数

基于Andun护心手表的血压测量主要依赖于四组PPG信号中的绿光PGG和三轴数据, 配合年龄、性别、身高、体重人体基本特征参数完成. 原始PPG信号和三轴信号数据如下图所示:

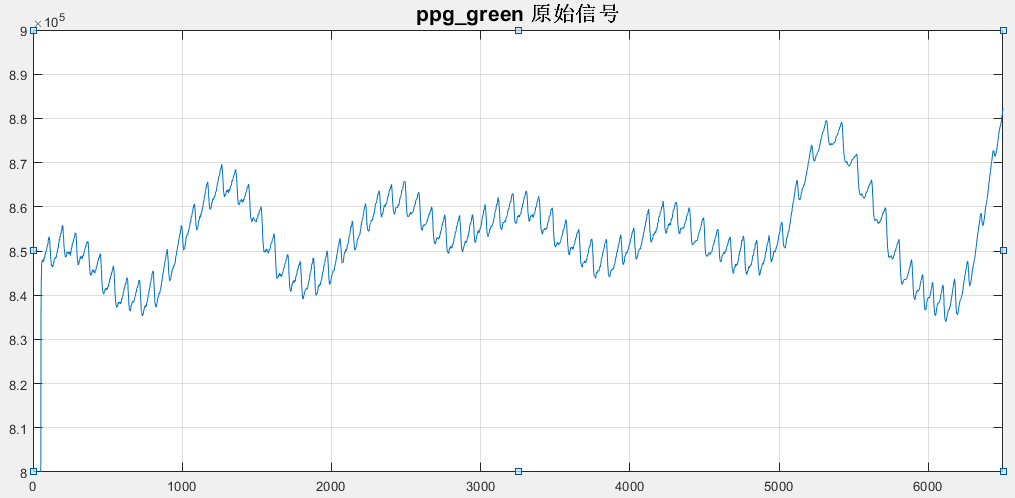


图. 原始绿光PPG信号数据

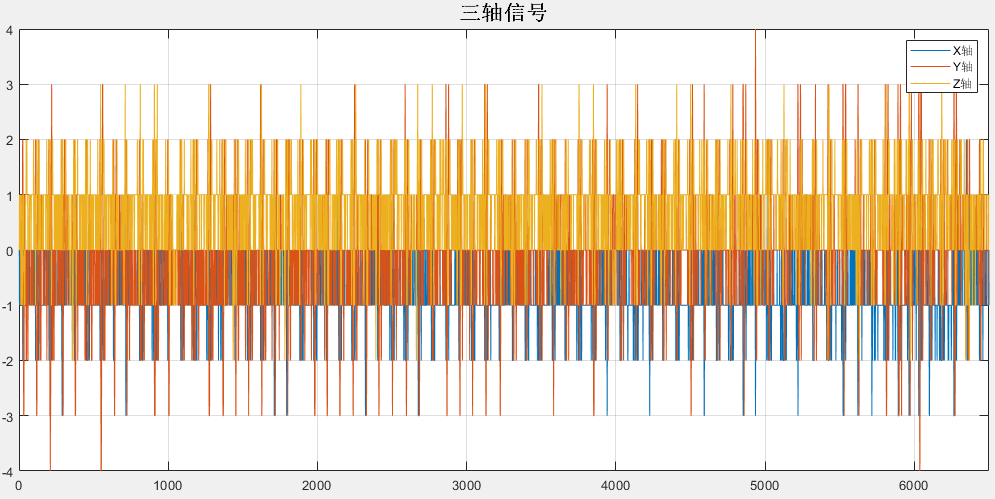


图. 原始三轴信号

在实际人体生理信号采集过程中, 由于信号微弱(通常为毫伏级)、信噪比低、检测时间长, 因此采集到的信号受到的干扰要比静态检测大得多, 这样再经过放大电路放大后的信号会经常淹没在较强的背景噪声中. 在采集过程中一般出现的噪声主要有高频噪声和基线漂移两种, 这些噪声主要是由仪器中电子元件、电磁场干扰、环境光的变化队及人体的运动等因素引起的. 高频和低频干扰会影响信号波形特征点提取的精确度, 因此有必要对信号波形进行去噪的处理.

PPG信号都属于低频人体生理信号, 频率范围在2~20Hz范围. 其高频干扰包括环境光、50Hz的工频等. 由于噪声的存在, PPG主波、重博波变得模糊不清. 工频噪声是由人体的分布电容以及50Hz电源频率引起的, 它的频率大约为50Hz或者是其倍频。由于这种由电源产生的噪声在硬件电路中无法消除，因此需要在软件中通过滤波算法进行滤除.

在PPG信号采集的过程中, 非常容易受到呼吸伪迹、运动伪迹的干扰. 在利用光电传感器采集PPG信号时, 信号容易受到外界光强的影响. 这些都将导致采集到的信号发生基线漂移.

#### 1.2 信号预处理

前面提到, PPG信号采集过程中会产生很大的噪声信号干扰, 因此在使用PPG计算血压之前需要对信号进行滤波处理. PPG噪声主要来源于高频的工频噪声和低频的基线漂移, 因此我们设计了高通和低通两个滤波器对PPG信号预处理.

* **高通滤波器:**

高通滤波为一个4阶的IIR滤波器, 其滤波器幅频相频特性图如下

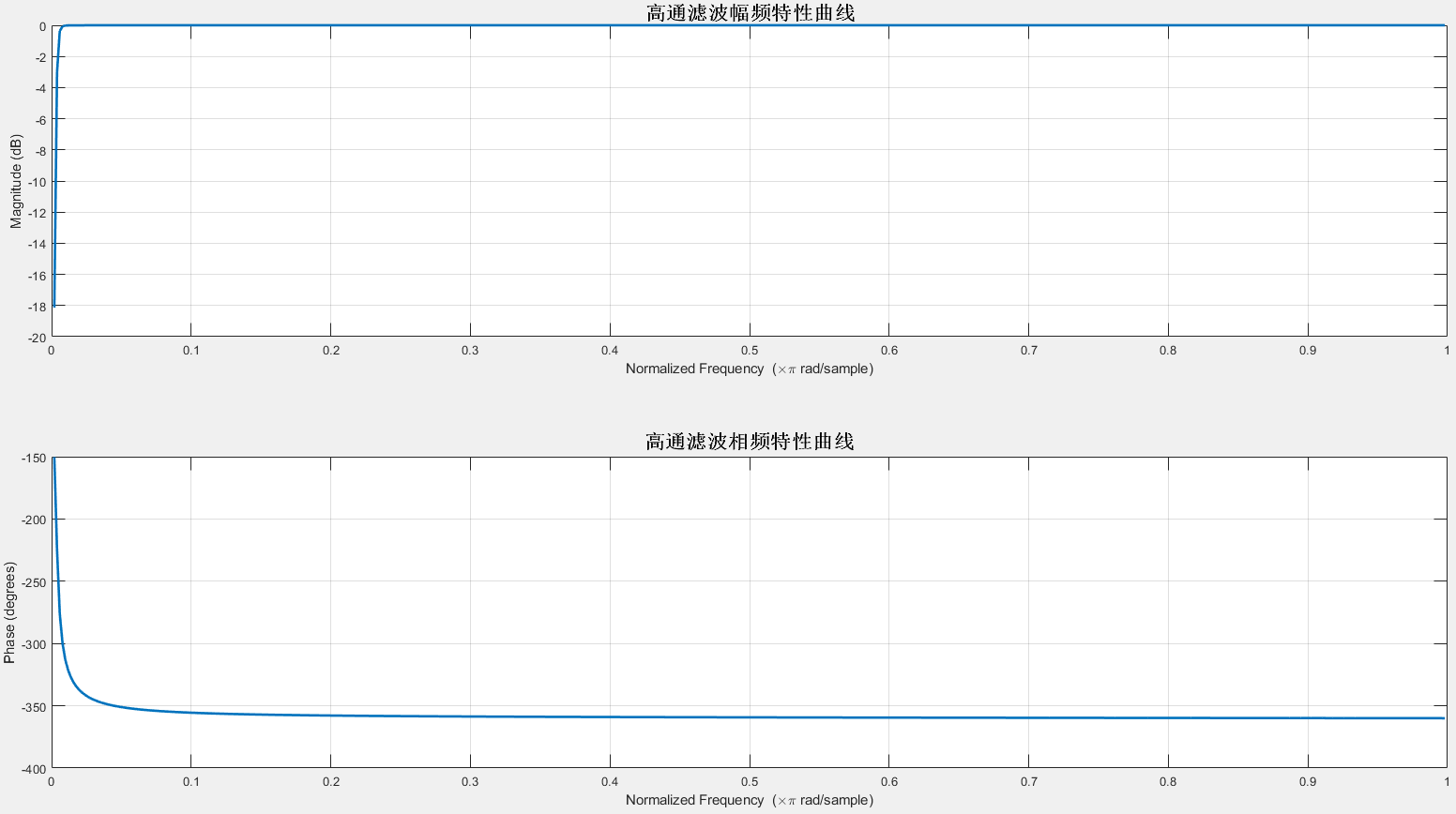


图. 高通滤波

* **低通滤波器:**

高通滤波为一个8阶的IIR滤波器, 其滤波器幅频相频特性图如下

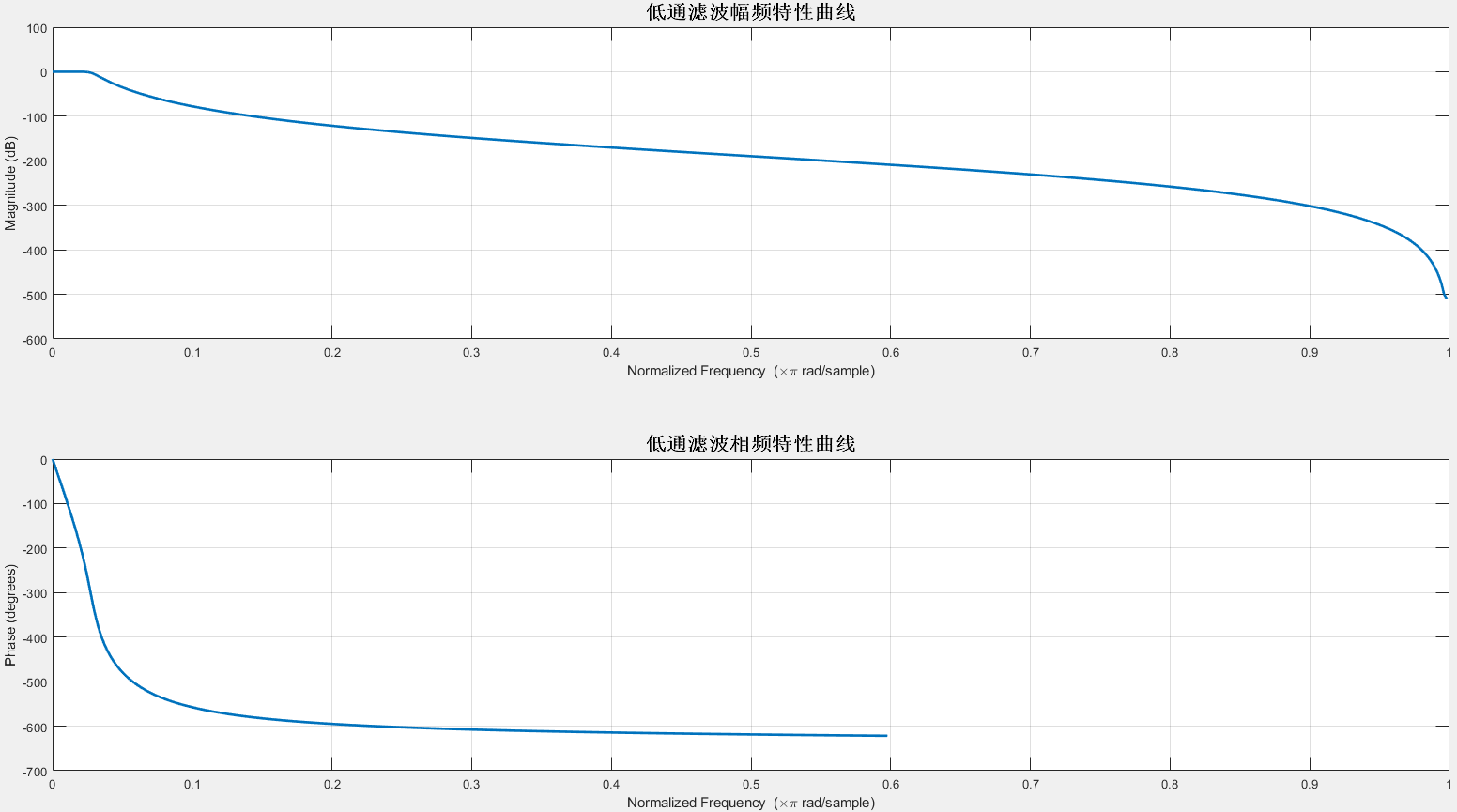


图. 低通滤波

* **滤波流程:**

#### 1.3 数据清洗

由于本次训练数据来源于两部分, 一部分为早期公司提供给UPI的数据, 另一部分为公司内部大规模采集的数据, 因此两部分数据的清洗清洗方案略有不同.

* **参考血压数据清洗**

对于早期公司提供提供给UPI的数据, 参考血压有PPG采集前和PPG采集后两个参考值, 在选取参考值时有如下规则:

1. 前后两次参考血压SBP差值不得超过16mmHg
2. 前后两次参考血压DBP差值不得超过12mmHg
3. 前后两次血压采集时对应的心率变化值不得超过8

除此之外, 两部分数据参考血压清洗规则基本相同

1. SBP最大值不得超过200, 最小值不得低于85
2. DBP最大值不得超过140, 最小值不得低于50
3. SBP不得小于或等于DBP
4. 某个血压值明显偏离于其他血压值

* **用户信息数据清洗**

在用户信息中, 主要删除了一些用户因体重单位明显错误的数据, 比如60kg错填为120kg.

* **PPG信号数据清洗**

PPG数据清洗主要是剔除一些采集过程中, 由于佩戴质量和运动等噪声数据, 对于是否运动使用三轴信号判断用户运动状态进行删除, 对于佩戴质量使用PPG\_Ambient信号辅助判断. 部分规则如下

1. 当AC特征低于300, 认为设备摘下
2. 当DSPLib参数返回为Reset标志, 认为运动状态
3. 当获取到的PPG特征每分钟少于15个, 认为运动状态

数据清洗从开始对接UPI开始就着手准备，主要工作包括调研、编码和实验，总共用时半个月完成。

### 2. 特征提取

#### 2.1 PPG血压估算参数

PPG特征由UPI直接提供, 我们负责调用相应相应接口获取特征即可, 下图是UPI提供的特征参数说明

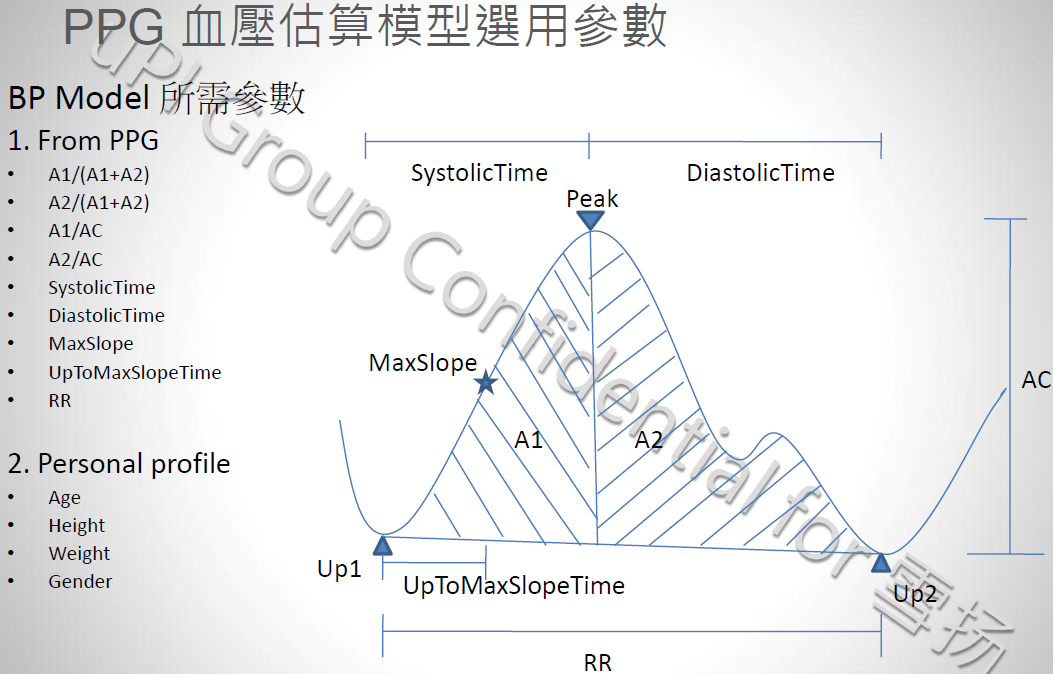


图. PPG特征说明

对于每一个PPG脉搏, 都会产生一组以上特征参数, 因此在实际应用应用中, 特征参数的产生速率和心率有关, 当心率高时产生的速度速度会变快变快反之亦然。

#### 2.2 特征提取

每一个脉搏都会产生一组特征参数, 我们并没有拿所有的特征参数参与模型训练, 在选取特征时选取16组特征中值作为一组训练训练数据, 同时每8个脉搏滑动产生下一组训练数据.

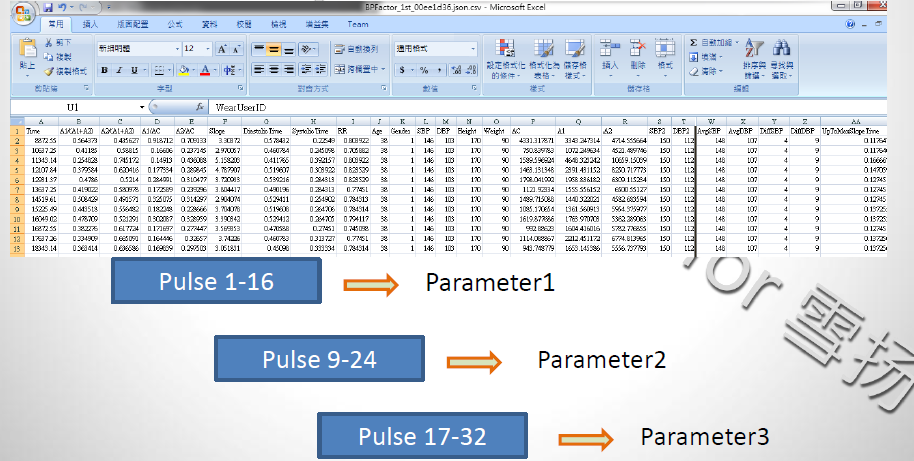


图. 特征提取方案说明

#### 2.3 线上PPG特征提取流程图

## 

## 算法与策略

### 模型结构

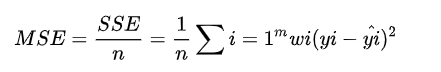
本次算法选用随机森林模型，随机森林是决策树模型的一种，决策树是一个[树形结构](https://link.zhihu.com/?target=https://baike.baidu.com/item/%E6%A0%91%E5%BD%A2%E7%BB%93%E6%9E%84" \t "_blank)。一个决策树在构建时，通过将数据划分为具有相似值的子集来构建出一个完整的树。决策树上每一个非叶节点都是一个特征属性的测试，经过每个特征属性的测试，会产生多个分支，而每个分支就是对于特征属性测试中某个值域的输出子集。决策树上每个叶子节点就是表达输出结果的连续或者离散的数据。

随机森林是以决策树为基础的一种更高级的算法，像决策树一样，随机森林既可以用户回归也可以用于分类。从名字中可以看出，随机森林是用随机的方式构建的一个森林，而这个森林是由很多的相互不关联的决策树组成。实时上随机森林从本质上属于机器学习的一个很重要的分支叫做集成学习。集成学习通过建立几个模型组合的来解决单一预测问题。它的工作原理是生成多个分类器/模型，各自独立地学习和作出预测。这些预测最后结合成单预测，因此优于任何一个单分类的做出预测。

所以理论上，随机森林的表现一般要优于单一的决策树，因为随机森林的结果是通过多个决策树结果投票来决定最后的结果。简单来说，随机森林中每个决策树都有一个自己的结果，随机森林通过统计每个决策树的结果，选择投票数最多的结果作为其最终结果。我觉得中国一句谚语很形象的表达了随机森林的运作模式，就是“三个臭皮匠，顶个诸葛亮”。随机森林算法具有更强的鲁棒性，对噪声数据不敏感。

### 损失函数

使用MSE作为训练的损失函数， 该统计参数是预测数据和原始数据对应点误差的平方和的均值，计算公式为：



其中，n为样本的个数。

### 正则化

计算方式为：IMG_256，防止模型过拟合。

### 模型融合

为了配合个人血压模型的预测，还使用随机森林的方法训练了另外2个通用模型，一个是使用2019年采集的血压数据训练的模型一，另一个是使用后台记录的用户校准的血压数据（2020-04月——2020-08月）进行训练得到的模型二，模型融合的方法如下：

1. 后台针对使用个人血压模型的用户单独建了一个数据表，表里面有3个字段old\_model\_weight,kefu\_model\_weight,personal\_model\_weight,分别代表模型一,模型二,和个人模型的权重;
2. 三个权重之和为1.0,一般默认个人模型的权重是1.0,其他两个通用模型的权重是0.0,意思就是默认只使用个人模型预测的血压值,当某个用户提供的血压数据较少,或者个人模型的训练误差偏大时,通过调整3个权重值的大小来条件最终的预测血压值;
3. 当手表上传了PPG特征数据时,分别使用如上3个模型进行预测,最终的血压值由每个模型的预测值乘以对应的权重加和得到;

### 出值策略

个人模型在输出血压值之前, 会根据上一次的血压值和时间, 对本次的输出值进行调整,调整的规则为: 如果上一次输出血压值的时间距离本次测量时间在30分钟以内, 则分别计算本次模型的高压值,低压值, 和上一次的高压值,低压值的变化幅度, 如果变化幅度超过20%, 则对本次的血压值按照20%的变化范围进行修正, 通过这种方式使模型计算的血压值不会在短时间内剧烈变化。

### 线上部署方式

线上依旧采用gRPC的实现方式，因为目前个人血压模型的用户较少,只部署了一个节点。

### 模型Retrain

用户加入新血压模型后,也会不定时的提供一些血压数据,对于这种情况,个人血压模型预留了接口,专门用于个人血压模型的再训练,使模型可以学习到用户新提供的血压数据特征。随着某个用户提供的血压数据越来越多，个人血压模型预测的也会越来越准确。

## 算法优化

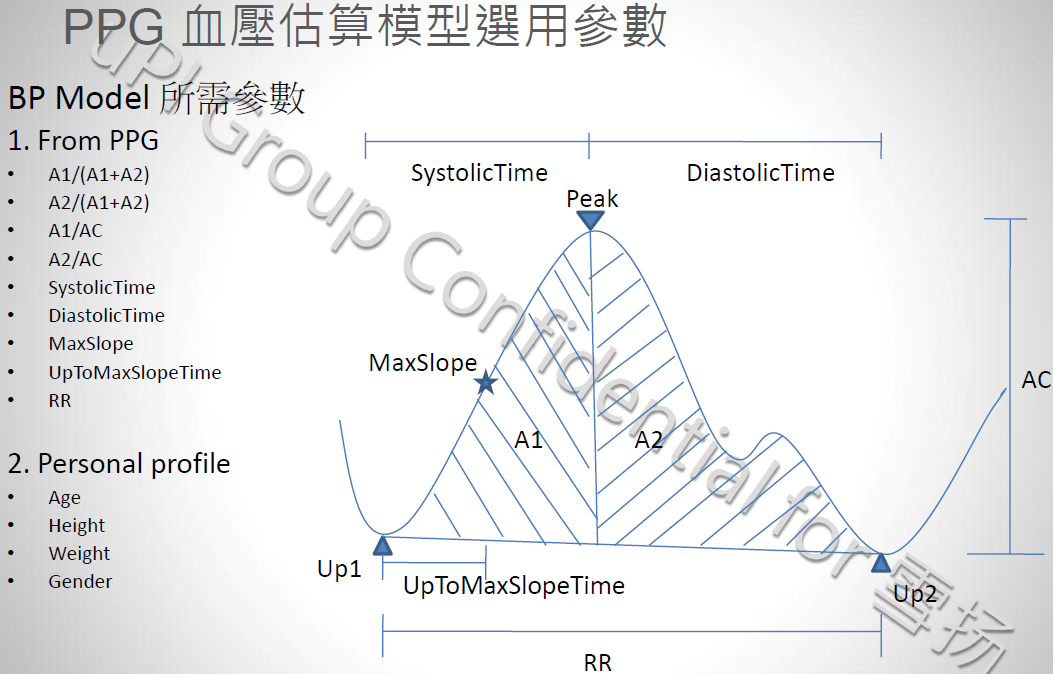
### 1. 基于算法模型的优化策略

当前血压计算方式是依赖于UPI对于PPG信号处理的基础上进行的, 因此我们的工作受限于UPI提供的特征. 在这一条件下, 提出以下算法模型优化策略.

#### 1.1 增加新的脉搏波波形特征

目前UPI的算法对一个脉动波形，提取了12个特征，实际上从波形提取的特征值只有8个特征，其他4个是由8个特征计算得到的，实际上从脉搏波波上还可以提取其他的波形特征，比如重搏波（如下图红色框）的幅值和时间，脉搏波波形90%峰值的持续时间等等，相关的一些论文研究表明，重搏波的变化是可以反映人体血管的硬化程度的。

具体需要增加什么特征，需要做试验，采集数据，通过比较观察来判断。比如可以在一个高血压用户身上采集其在高压，低压时候的原始脉搏波波形，通过人工观察波形的变化点，来确定增加的特征点。



#### 1.3 PPG特征值的异常检测

异常值检查是专门的一套算法，是通过统计规律判断数值的分布规律。对于血压预测来说，通过分析用户过去佩戴历史的PPG特征数值的分布规律,对新产生的PPG特征值进行判断, 查看此时测量的PPG特征值是否符合历史规律,或者说是否是之前经常出现的,如果是,则说明此用户当前的血压正常,如果此时的ppg特征值是之前很少出现,或者根本没有出现过的,那说明此时的血压不正常(可能是血压过高,或者过低)。

此算法的好处是不需要用户提供血压值，只是通过PPG特征参数进行判断；但是此算法依赖于PPG特征参数计算的准确性，及PPG特征参数能否反映血压的变化的前提下，才有意义。

## 当前算法存在的问题

对于当前算法适应于实际场景，还有如下方面需要改善：

1. 个人模型依赖于用户测量的血压数据，而且用户提供的血压数据要能匹配到手表测量的PPG特征数据；
2. 目前的PPG特征参数是12个，但是这个12个参数是否能真实反映血压的变化，还没有严格的试验验证，接下来需要制定计划，对PPG波形变化和血压变化直接的关系做试验，收集试验数据；
3. 对于特征提取方面没有、或者有很局限的优化空间。因为传感器提供商upi提供的特征方法是对外保密，现有技术储备还不能超越之前的特征，所以对特征提取方面的技术有短板。

## 项目未来的发展方向

### 工程实现

对于提供血压数据较多，需要进行血压优化的用户，现在的方案是每天手动筛查，把满足条件的用户加入到个人血压模型，而且对于之前加入个人血压模型，又提供新数据的用户，需要使用新的血压数据进行模型的再训练（retrain）。将来可以以脚本或者接口的形式，检测用户提供的血压数据的变化，自动调整模型，从而更及时的对用户血压模型进行调整。

### 硬件角度

考虑PPG信号的局限性，和对动态时监测的极不稳定性质。在有条件的情况下，加入ECG等其他传感器，获取更多生命体征信号，对于分析心脑血管疾病，以及测量心率、血压、血氧、心电，都有着重要帮助。