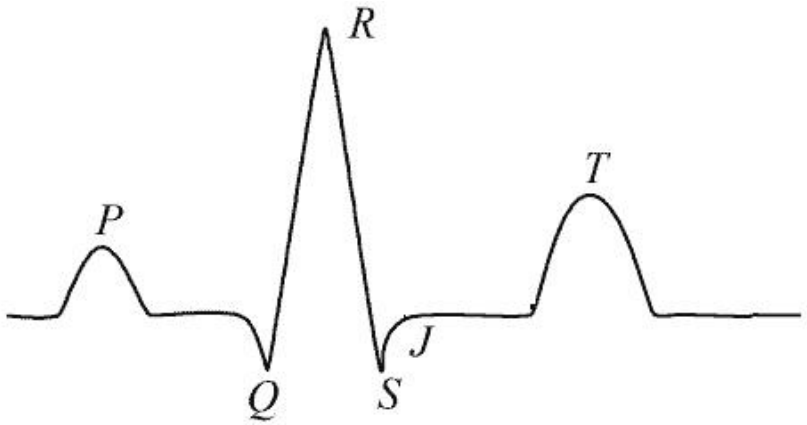
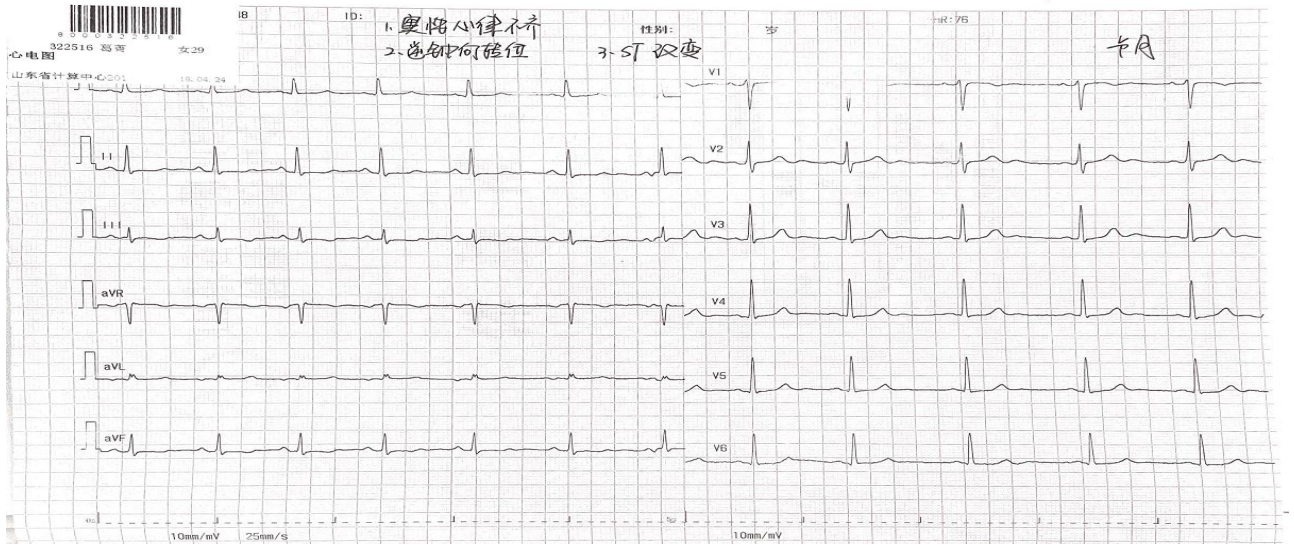
基于集成学习的心电数据智能分类

2022.11.9-10 李娜 王迪

ECG简介

- 心电图（ECG）是利用心电图机从体表记录心脏每一心动周期所产生的电活动变化图形的技术





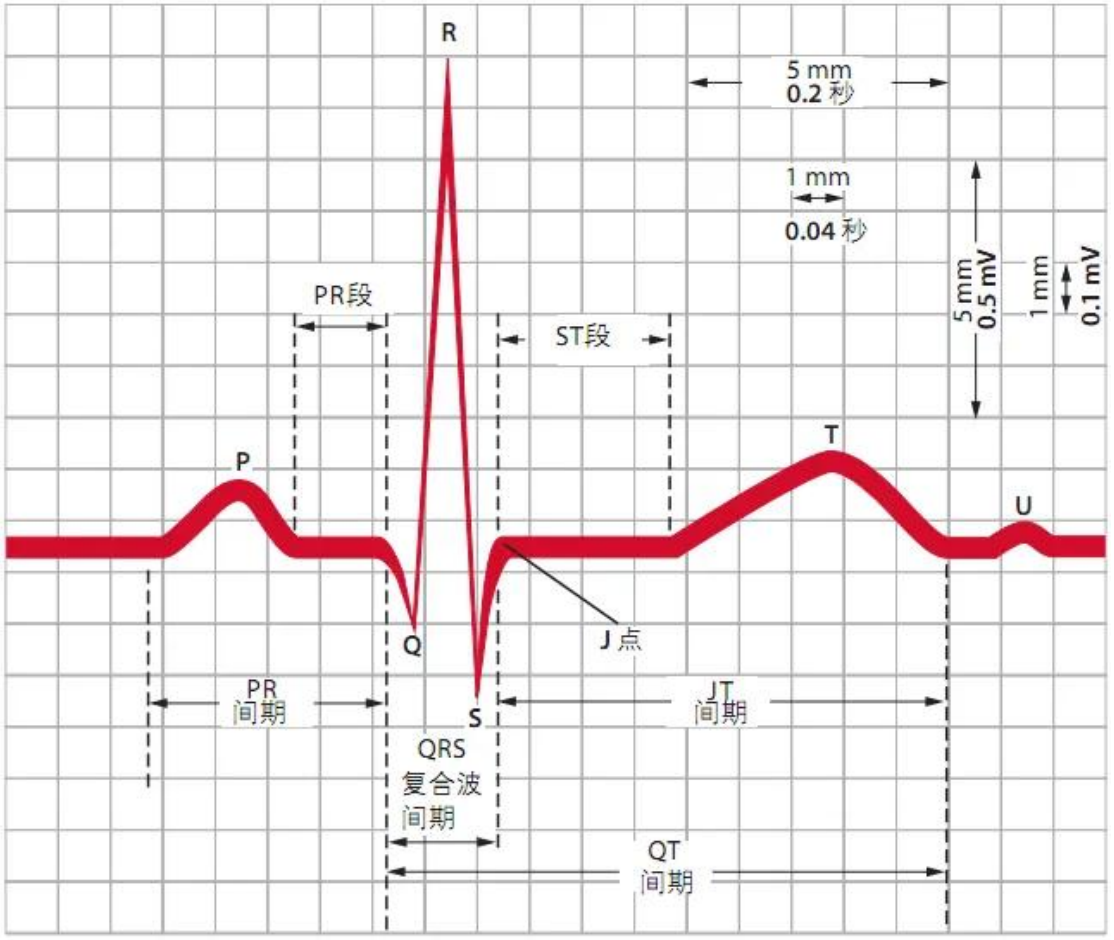
- 一个完整周期的ECG信号有 QRS波、P波、T波组成，不同的人对应不用的波形，同一个人在不同的阶段波形也不同。我们须要依据各个波形的特点，提取出相应的特征，从而对心电图进行分类诊断。

案例背景

- 随着全球人口老龄化问题的日益加剧，患心脏疾病的人群日益增加。据不完全统计，全世界死亡人口中大约有三分之一属于心脏疾病；在我国，每年也有大约54万人死于心脏疾病。
- 心脏疾病及其引发的其他心血管疾病正不断威胁着人类健康，通过各种方式提前预防、诊断心血管疾病显得尤为重要。
- 随着穿戴式心电设备的普及，心电图的获取日益简单，但由于只有专业医师才能解读心电图，严重制约着心电图的应用。

案例背景

- 心电图自动分析应运而生，其融合了传感器技术、信号处理技术、模式识别技术等，是医学和信息技术交叉领域最典型的应用点之一。
- 研究智能模型，实现心电图的智能诊断，从而使普通人也能看懂心电图。



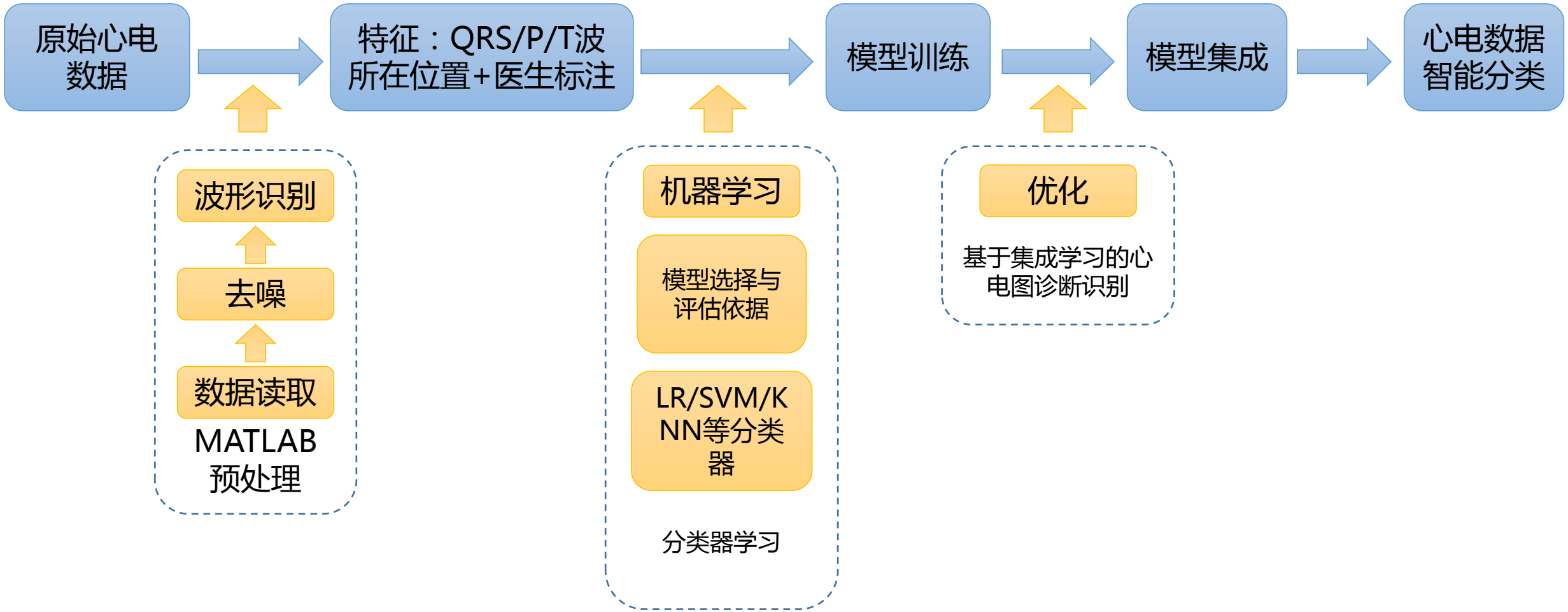
案例目的

- 针对心电数据进行基于小波变换的去噪，并实现QRS、P、T波形识别，获得QRS、P、T波形的位置点以及医生的标注数据
- 针对心电数据进行“正常、房颤、房性早搏、偶发房性早搏、频发房性早搏、房性心动过速、房颤伴快速心室率”，这七种诊断的智能识别分类。

案例任务

- 编写MATLAB程序，实现心电数据的预处理，获得心电数据的特征
- 了解机器学习的基本原理
- 编写Python脚本，设计机器学习算法模型，进行心电数据的模型训练
- 编写Python脚本，实现算法模型的集成。

案例流程


原始心电数据
特征：QRS/P/T波所在位置+医生标注
模型训练
模型集成
心电数据智能分类
波形识别
去噪
数据读取
MATLAB预处理
机器学习
模型选择与评估依据
LR/SVM/KNN等分类器
分类器学习
优化
基于集成学习的心电图诊断识别

实验步骤概述

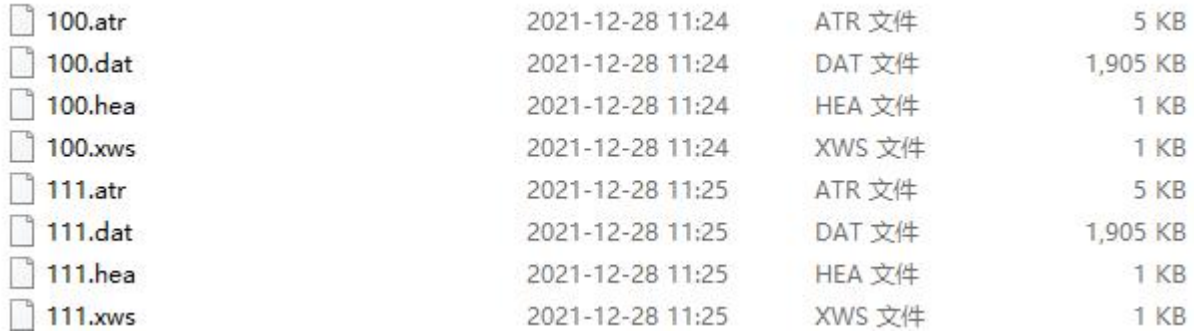
- 本案例共包括3个实验步骤
  - 利用MATLAB实现心电数据的预处理
  - 设计机器学习模型，利用Python实现模型训练
  - 集成模型，获得更为准确的分类模型

MIT-BIH数据库简介

- MIT-BIH 是由美国麻省理工学院提供的研究心律失常的数据库，是目前国际公认的三大心电数据库之一，其包含了48组经过注释的心率失常心电记录。
- 该数据库包括47个测试个体的4000多个24小时的周期性动态心电数据，有48个时长约为30min的记录文件，共计109500个心拍，其中异常心拍约占30%。

数据格式

- MIT-BIH 为了节省文件长度和存储空间，使用了自定义的格式。一个心电记录由三个部分组成：
  - 头文件[.hea]，存储方式ASCII码字符，对数据文件进行格式说明
  - 数据文件[.dat]，按二进制存储，每三个字节存储两个数，一个数12bit，是具体的心电信号的ADC转换值。
  - 注释文件[.atr]，按二进制存储，为心电专家的诊断信息。



100.atr	2021-12-28 11:24	ATR 文件	5 KB
100.dat	2021-12-28 11:24	DAT 文件	1,905 KB
100.hea	2021-12-28 11:24	HEA 文件	1 KB
100.xws	2021-12-28 11:24	XWS 文件	1 KB
111.atr	2021-12-28 11:25	ATR 文件	5 KB
111.dat	2021-12-28 11:25	DAT 文件	1,905 KB
111.hea	2021-12-28 11:25	HEA 文件	1 KB
111.xws	2021-12-28 11:25	XWS 文件	1 KB

[.hea]头文件

- [.hea]为头文件，其由一行或多行ASCII码字符组成。以111.hea为例

第一行为记录行，指出该记录为采样率为360Hz、采样点数为650000的两路导联信号

111.dat 212 200 11 1024 1017 20838 0 MLII

111.dat 212 200 11 1024 1031 12452 0 V1

为信号技术规范说明行，表示信号以12位的位压缩格式（即“212”格式）进行存储的，信号增益都是每200ADC uints/mV，电压模数转换ADC的分辨率为11位，ADC的转换为1024代表0V。两个信号的第一采样点的ADC值分别为1017和1031，采样点的校验数分别为20838和12452，0表示输入输出可以以任何尺寸的块，两个信号分别采自MLII导联和V1导联

47 F 937 167 x1

Digoxin, Lasix

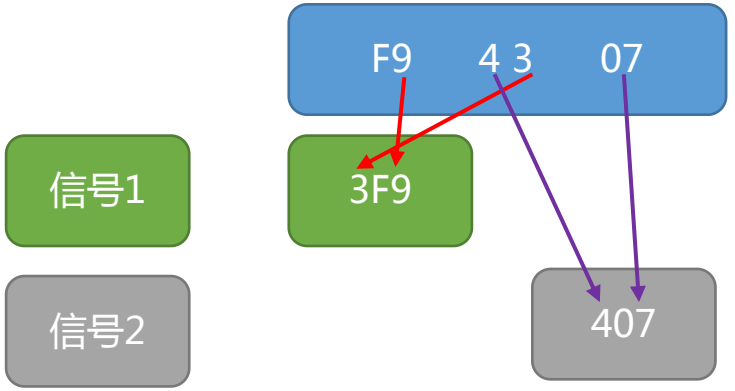
文件的最后两行包含了注释字符串，其中第一行说明了患者的性别和年龄以及记录数据，第二行列出了患者的用药情况。

*MIT—BIH心率失常数据库的识读及应用，宋喜国，中国医学物理学杂志，2004年7月第21卷第4期

[.dat]数据文件

- 心律失常数据库统一采用212格式进行存储
- “212”格式是针对两个信号的数据库记录，这两个信号的数据交替存储，每三个字节存储两个数据。
- 假设这两个数据分别采样自信号1和信号2
  - 信号1的采样数据取自第一、二字节(16位)的12位，其中第一字节作为低8位，第二字节的低4位作为其高4位；
  - 信号2的采样数据由第二字节的高4位（作为组成信号1采样数据的12位的高4位）和下一字节的8位（作为组成信号1采样数据的12位的低8位）共同组成。

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00000000	F9	43	07	F9	43	07	F9	43	07	F9	43	07	F9	43	07	F9
00000010	43	07	F9	43	07	F9	43	07	F9	43	07	F9	43	0C	FC	43
00000020	11	F8	43	0E	F3	43	09	F0	43	02	EC	43	02	EE	43	03
00000030	F0	43	01	EB	43	00	ED	33	FB	E7	33	F5	E5	33	F7	E7
00000040	33	F8	E9	33	FB	EB	33	FB	E8	33	FB	E6	33	FA	E7	33



[.dat]数据文件

- 用MATLAB移位实现信号1与信号2
  - 以第一组为“F9 43 07”为例
  - HEX: F9 43 07
  - DEC: 249 67 07
  - BIN: 1111 1001 | 0100 0011 | 0000 0111

```
%对第二字节做右位移运算，位移距离4
%得到第二字节左四位，即sign2的高四位
M2H = bitshift(A(:,2), -4);
%对第二字节和00001111做与运算，
%保留第二字节右四位，即sign1的低四位
M1H = bitand(A(:,2), 15);
```

A | M1H | M2H

10000x3 double

	1	2	3
1	249	67	7
2	249	67	7
3	249	67	7

67(DEC):01000011(BIN)

右移四位→0000 0100(BIN)→4(DEC)

M2H

67(DEC):01000011(BIN)

与运算→0100 0011(BIN)
0000 1111
→0000 0011(BIN)
→3(DEC)

M1H

A | M1H | M2H

10000x1 double

	1	2
1	3	
2	3	
3	3	

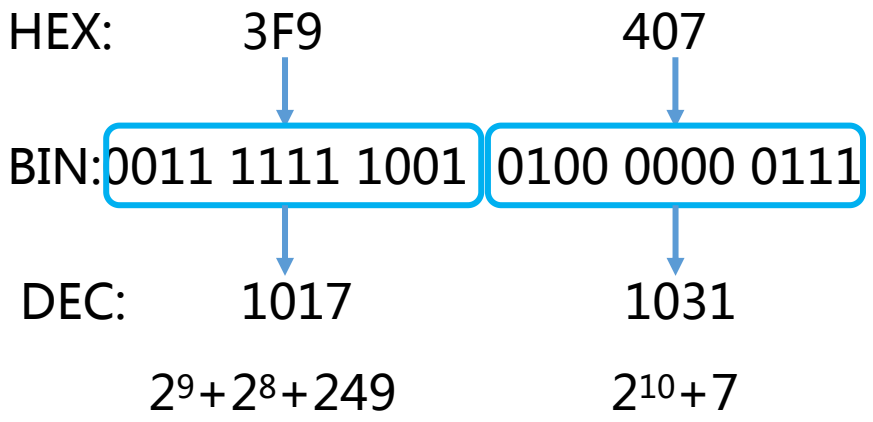
A | M1H | M2H

10000x1 double

	1	2
1	4	
2	4	
3	4	

[.dat]数据文件

- 以第一组为“F9 43 07”为例
  - 两个值则分别为0x3F9和0x407
  - 转换为十进制分别为1017和1031
  - 单位是mv，增益为200
  - 0点为1024mv
  - 所以代表的信号幅度分别为
    - (1017mv-1024mv)÷200=-0.035mv
    - (1031 mv-1024mv)÷200=0.035mv
  - 这两个值分别是两个信号的第一采样点对应的幅值，后面依此类推，分别表示了两个信号的采样幅值。




```
M1H = bitand(A(:,2), 15);
%对第二字节和00001000做与运算，
%保留第二字节右边第四位，获取sign2符号位,并向左位移九位，与整体sign1进行运算
PRL=bitshift(bitand(A(:,2),8),9);
%对第二字节和10000000做与运算，
%保留第二字节右边第四位，获取sign1符号位，并向左位移5位，与整体sign2进行运算
PRR=bitshift(bitand(A(:,2),128),5);

M( : , 1)= bitshift(M1H,8)+ A(:,1)-PRL;
M( : , 2)= bitshift(M2H,8)+ A(:,3)-PRR;
```

```
if M(1,:) ~= firstvalue
    error('inconsistency in the first bit values');
end
switch nosig
case 2
    M( : , 1)= (M( : , 1)- zerovalue(1))/gain(1);
    M( : , 2)= (M( : , 2)- zerovalue(2))/gain(2);
    TIME=(0:(SAMPLES2READ-1))/sfreq;
```

[.atr]注释文件

- 记录了心电专家对相应的心电信号的诊断信息，心律失常数据库采用的MIT格式进行数据标注
- MIT格式是一种紧凑型格式，每一注释的长度占用偶数个字节空间，多数情况下是占用两个字节，多用于在线的注释文件
- MIT格式，每一注释单元的前两个字节的第一个字节为最低有效位，16位中的最高6位表示了注释类型代码，剩余的10位说明了该注释点的发生时间或辅助信息，若为发生时间，其值为该注释点到前一注释点的间隔（对于第一个注释点为从记录开始到该点的间隔），若为辅助信息则说明了附加信息的长度。

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00000000	1F	70	03	FC	28	4E	00	00	A6	08	24	09	3B	09	43	09
00000010	2F	09	35	09	2D	09	30	09	29	09	40	09	3D	09	2D	09
00000020	34	09	34	09	2B	09	24	09	3B	09	3E	09	32	09	3F	09
00000030	2B	09	27	09	28	09	47	09	3D	09	32	09	37	09	30	09

0x701F
0xFC03

[.atr]注释文件

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00000000	1F	70	03	FC	28	4E	00	00	A6	08	24	09	3B	09	43	09
00000010	2F	09	35	09	2D	09	30	09	29	09	40	09	3D	09	2D	09
00000020	34	09	34	09	2B	09	24	09	3B	09	3E	09	32	09	3F	09
00000030	2B	09	27	09	28	09	47	09	3D	09	32	09	37	09	30	09

- 16位值0x701F(0111 0000 0001 1111)
  - 高6位的值为0111 00(BIN)(十进制28)，该类型代码为28，代表意义是节律变化
  - 低10位的值为00 0001 1111(BIN)(十进制31)，发生时间在0.086秒(31/360Hz)

```
bitshift(A(i,2),-2);%右移2位
```

```
bitshift(bitand(A(i,2),3),8)+A(i,1)
```

变量 - A

A

2142x2 double

	1	2	3	4
1	31	112		
2	3	252		
3	40	78		
4	0	0		
5	166	8		

- 16位值0xFC03(1111 1100 0000 0011)
  - 高6位的值为0x3F(1111 11(BIN))(十进制63)，该类型代码为63，代表的意义是在该16位值后附加了3个（低10位的值为0x03，低10位值代表的数）字节的辅助信息，若字节个数为奇数，则再附加一个字节的空值，在本例中就是“28 4E 00 00”
- 16位值0x08A6(0000 1000 1010 0110)
  - 高6位的值为0000 10(BIN)(十进制2)，该类型码2代表左束支传导阻滞
  - 低10位的值为00 1010 0110(BIN)0xA6(十进制166)，发生时间为0.547秒((31+166)/360Hz)
- 当高6位为十进制59时，读取之后第3个16位的高6位，作为类型代码，读取之后第二个16位+第一个16位*2^16；
- 高6位为十进制60，61，62时，继续读下一个16位。

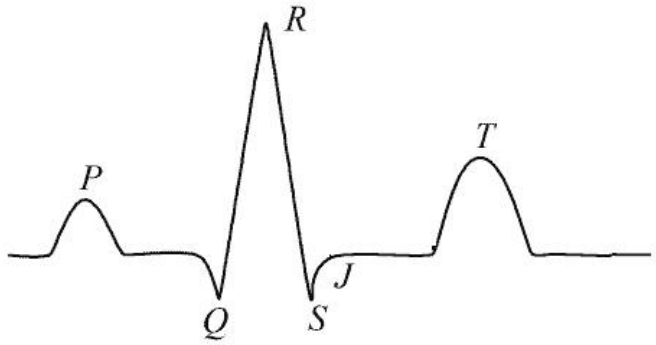
数据去噪

- 原始ECG信号含有高频噪声与基线漂移
- 利用小波分解去除对于噪声
  - 采用8层小波分解，小波类型为双正交小波bior2.6
  - 1、2层细节系数即高频信息置0
  - 8层的近似系数即低频漂移置0
- 小波重构去噪

QRS波检测

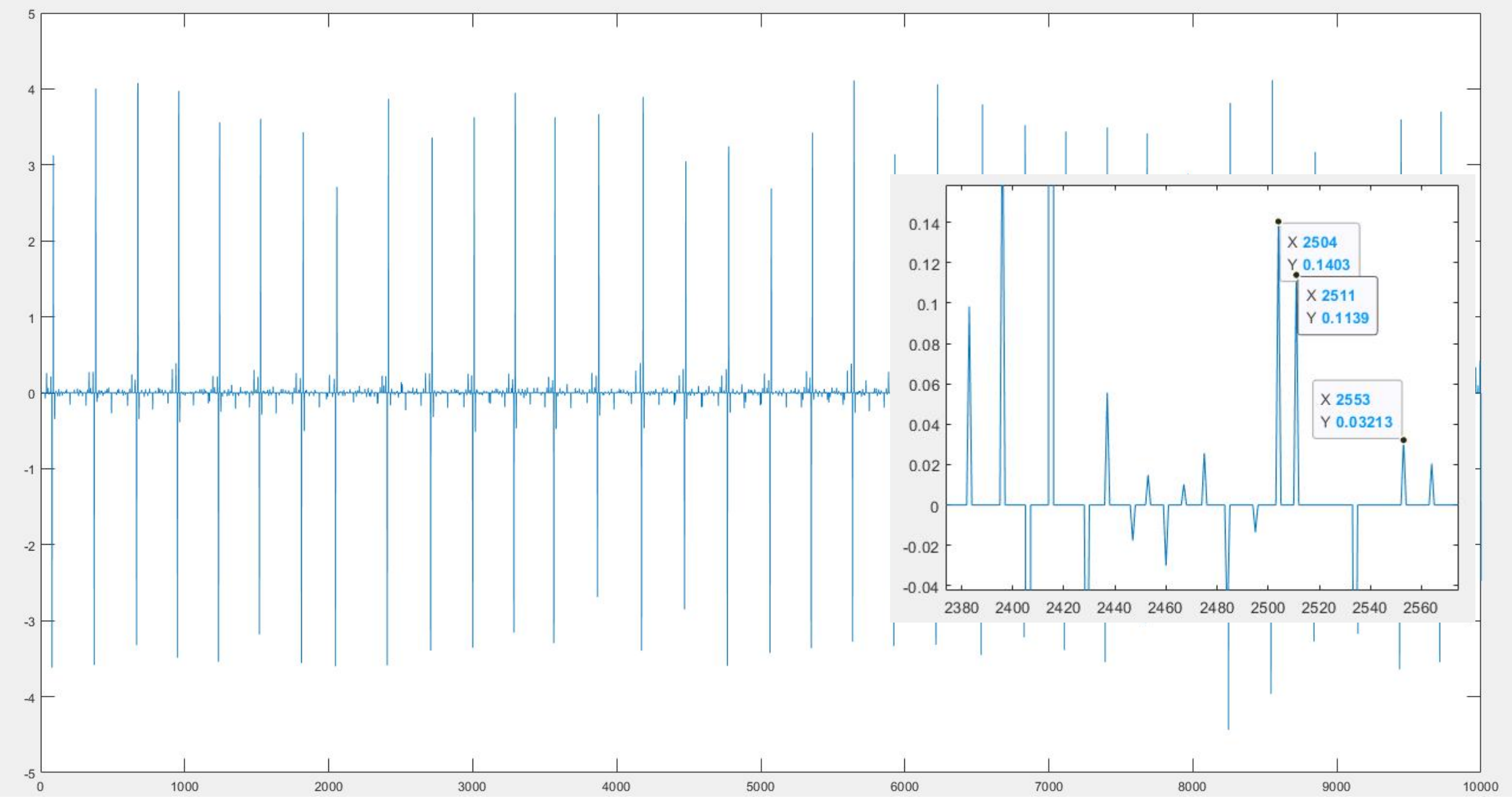
- QRS检测是处理ECG信号的基础，不管最后实现什么样的功能，QRS波的检测都是前提。
- 采用基于二进样条4层小波变换。在3层的细节系数中利用极大极小值方法能够非常好的检测出R波。3层细节系数的选择是基于R波在3层系数下表现的与其它噪声区别最大。

R波检测



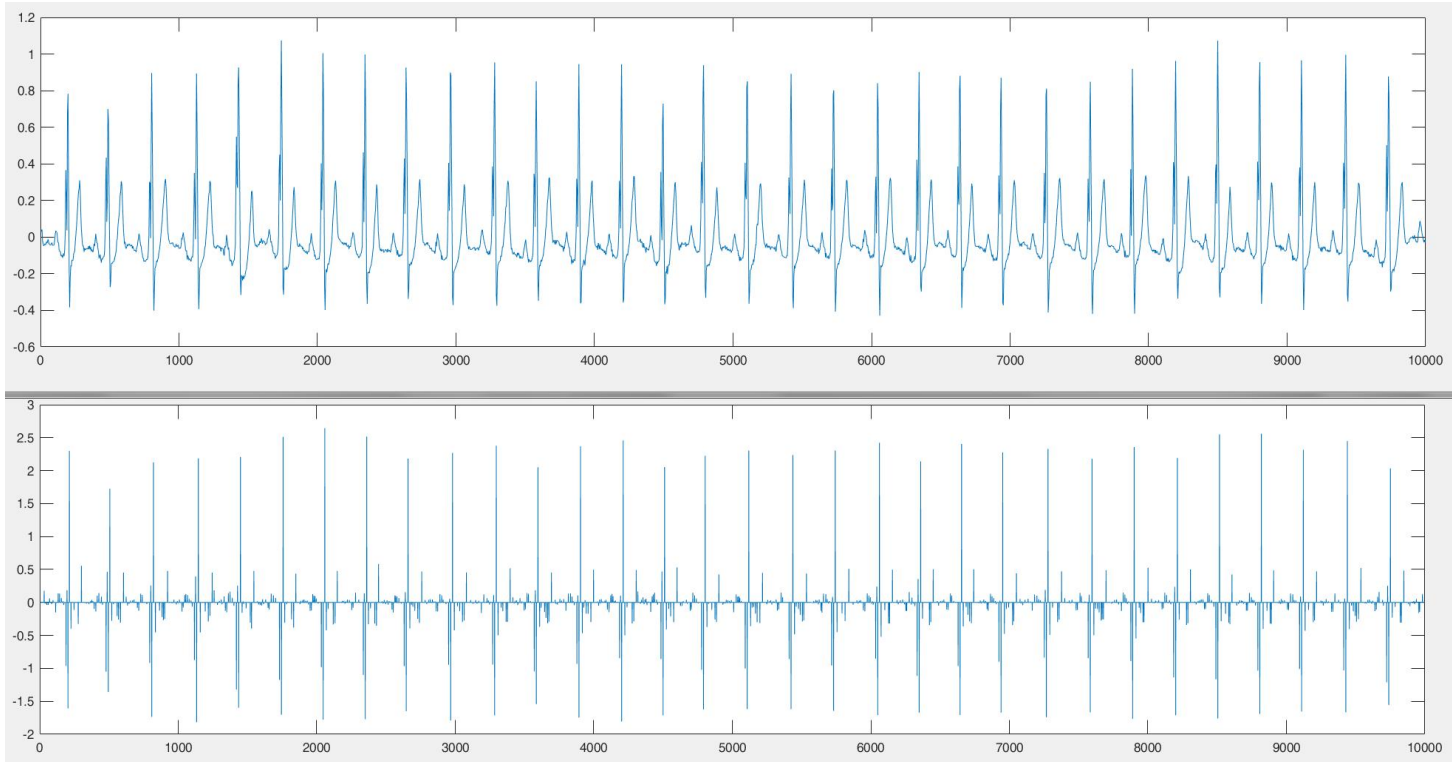
- 3层细节系数找出极大极小值对：
  - 极大值：找出斜率>0的位置点，然后在这些位置点中找到前面一个位置对应的值比后面的点的值大的位置点，并赋值为1，其余为0
  - 极小值：找出斜率<0的位置点，然后在这些位置点中找到前面一个位置对应的值比后面的点的值大的位置点，并赋值为1，其余为0

```
%小波系数的大于0的点
posw=swd.*(swd>0);
%斜率大于0
pdw=((posw(:,1:points-1)-posw(:,2:points))<0);
%正极大值点
pddw(:,2:points-1)=((pdw(:,1:points-2)-pdw(:,2:points-1))>0);
%小波系数小于0的点
negw=swd.*(swd<0);
%斜率小于0
ndw=((negw(:,1:points-1)-negw(:,2:points))>0);
%负极大值点
nddw(:,2:points-1)=((ndw(:,1:points-2)-ndw(:,2:points-1))>0);
```

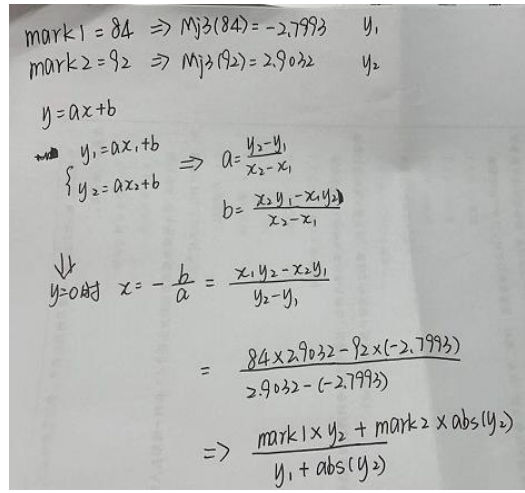

X 2504
Y 0.1403
X 2511
Y 0.1139
X 2553
Y 0.03213

R波检测

- 设置阈值，提取R波
  - R波的值明显高于其他位置的值，在3层细节系数的特点也是如此
  - 设置可靠阈值，提取一组相邻的最大最小值对，这个最大最小值对的过0点对应与原始信号的R波点



```
while i<points
    if interva2(i)==-1
        mark1=i;
        i=i+1;
        while(i<points&interva2(i)==0)
            i=i+1;
        end
        mark2=i;
%求极大值对的过零点
        mark3= round((abs(Mj3(mark2))*mark1+mark2*abs(Mj3(mark1)))/(abs(Mj3(mark2))+abs(Mj3(mark1))))
%R波极大值点
        R_result(j)=mark3;
        countR(mark3)=1;
```



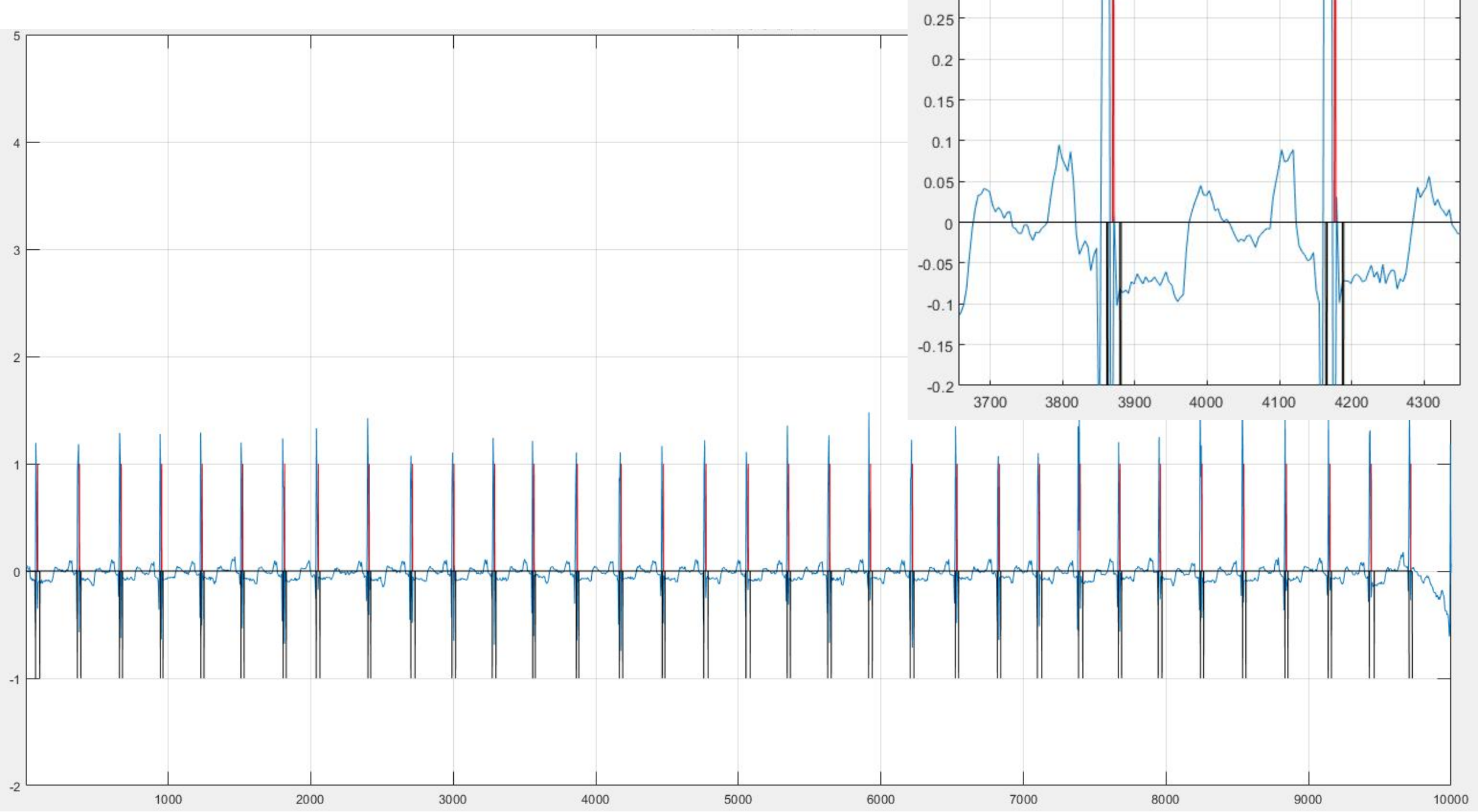
Q/S波检测

- 找出Q波和S波
  - 基于R波的位置
    - 在R波位置（1层细节系数下）的前3个极点为Q波
    - 在R波位置的后三个极点为S波

```
%R波极大值点
        R_result(j)=mark3;
        countR(mark3)=1;
%求出QRS波起点
        kqs=mark3;
        markq=0;
      while (kqs>1)&&( markq< 3)
          if Mj1(kqs)~=0
             markq=markq+1;
          end
          kqs= kqs -1;
      end
   countQ(kqs)=-1;
%求出QRS波终点
   kqs=mark3;%-10
   marks=0;
   while (kqs<points)&&( marks<3)
       if Mj1(kqs)~=0
          marks=marks+1;
       end
       kqs= kqs+1;
   end
   countS(kqs)=-1;
```

QRS波检测

- R波漏检与错检
  - 错检
    - T波检测为R波
    - 相邻R波距离<0.4mean(RR)
    - 删除
  - 漏检
    - 相邻R波距离>1.6mean(RR)
    - 两个R波之间找到一个最大的极值对
    - 添加R波


5
4
3
2
1
0
-1
-2
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
3700
3800
3900
4000
4100
4200
4300

P波与T波检测

- P波与T波在4层细节系数中可以表述出更好的特性，同样根据极大极小值原理，可以分别检测出T波与P波的位置以及对应的起始点与终点。