

数据挖掘作业一

数据分析过程报告

（马的疝病分析）

姓名：董国昭

学院：生命学院

学号：2120161295

一、 题目说明

马的疝病分析

1. 问题描述

疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。

2. 数据说明

共 368 个样本，27 个特征。关于特征的详细说明见下载链接。

3. 数据分析要求

3.1 数据可视化和摘要

数据摘要

对标称属性，给出每个可能取值的频数，
数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。
数据的可视化

针对数值属性，

绘制直方图，用 qq 图检验其分布是否为正态分布。
绘制盒图，对离群值进行识别。

3.2 数据缺失的处理

数据集中有 30% 的值是缺失的，因此需要先处理数据中的缺失值。

分别使用下列四种策略对缺失值进行处理：

将缺失部分剔除

用最高频率值来填补缺失值

通过属性的相关关系来填补缺失值

通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

二、 数据整合及调整

基于 Python 进行数据的初步整合，对原始数据进行调整得到了数据集为 28*368 的“.csv”文件，通过 Excel 重新将数据类型改变为“.xlsx”类型的文件。

利用 MATLAB 平台进行相关的后续处理。

通过对说明文件进行分析可知，属于对标称属性的数据有 20 个，其编号为“1,2,7,8,9,10,11,12,13,14,15,17,18,21,23,24,25,26,27,28”，属于数值属性的数据有 7 个，其编号为“4,5,6,16,19,20,22”。其中第 3 个为 ID 号，不作为数据指标进行分析。

三、 数据可视化和摘要

1. 算法流程及处理过程

- 1) 首先利用 MATLAB 进行数据的读取，保存为组别（28）*样本数（368）的二维矩阵，其中缺失值被定义为 NAN，并且完成组别名称对应。
- 2) 筛选 20 个对标称属性的数据集，利用 MATLAB 中的 tabulate 函数进行频数的统计，并输出相关的统计结果。
- 3) 筛选出 7 个数值属性的数据集，利用 max 函数求最大值，min 函数求最小值，mean 函数求取均值，prctile 函数求取中位数及四分位数，isnan 函数求取缺失值的数量，并输出相关的统计结果。
- 4) 通过 hist 函数，qqplot 函数，boxplot 函数进行数值属性数据的直方图，Q-Q 图以及盒图的描绘，输出相关的结果图。

2. 结果输出展示

1) 20 个对标称属性数据集的取值及对应频数的统计结果

| 数据名称 | 数值 | 频数 | 数据名称 | 数值 | 频数 | 数据名称 | 数值 | 频数 |
|----------------------------|----|-----|------------------|----|-----|----------------------|----|-----|
| surgery | 1 | 214 | mucous membranes | 1 | 98 | abdominal distension | 1 | 101 |
| | 2 | 152 | | 2 | 38 | | 2 | 75 |
| age | 1 | 340 | | 3 | 81 | | 3 | 82 |
| | 9 | 28 | | 4 | 50 | | 4 | 42 |
| temperature of extremities | 1 | 95 | | 5 | 28 | nasogastric tube | 1 | 89 |
| | 2 | 39 | | 6 | 25 | | 2 | 121 |
| | 3 | 135 | pain | 1 | 49 | | 3 | 27 |
| | 4 | 34 | | 2 | 77 | nasogastric reflux | 1 | 141 |
| peripheral pulse | 1 | 151 | | 3 | 82 | | 2 | 45 |
| | 2 | 6 | | 4 | 47 | | 3 | 49 |
| | 3 | 116 | | 5 | 50 | abdomen | 1 | 31 |
| | 4 | 12 | peristalsis | 1 | 49 | | 2 | 24 |
| capillary refill time | 1 | 232 | | 2 | 22 | | 3 | 19 |
| | 2 | 96 | | 3 | 154 | | 4 | 55 |
| | 3 | 2 | | 4 | 91 | | 5 | 96 |

| 数据名称 | 数值 | 频数 | 数据名称 | 数值 | 频数 | 数据名称 | 数值 | 频数 |
|--------------------|----|-----|-----------------------------|------|-----|-------------------|------|-----|
| rectal examination | 1 | 68 | abdominocentesis appearance | 1 | 52 | type of lesion 26 | 0 | 358 |
| | 2 | 14 | | 2 | 62 | | 1400 | 1 |
| | 3 | 61 | | 3 | 60 | | 2208 | 11 |
| | 4 | 97 | outcome | 1 | 225 | | 3111 | 3 |
| cp_data | 1 | 124 | | 2 | 89 | | 3112 | 1 |
| | 2 | 244 | | 3 | 52 | | 3205 | 2 |
| surgical lesion | 1 | 232 | type of lesion 27 | 0 | 367 | | 6112 | 1 |
| | 2 | 136 | | 2209 | 1 | | 7111 | 1 |

数据名称 type of lesion level（因元素较多，故未收入表格）

'in_type of lesion level'

the number of _0 is _67
the number of _300 is _1
the number of _400 is _7
the number of _1111 is _1
the number of _1124 is _2
the number of _1400 is _10
the number of _2111 is _4
the number of _2112 is _6
the number of _2113 is _8
the number of _2124 is _9
the number of _2205 is _17
the number of _2206 is _5
the number of _2207 is _3
the number of _2208 is _23
the number of _2209 is _15
the number of _2300 is _2
the number of _2305 is _1
the number of _2322 is _2
the number of _3025 is _2
the number of _3111 is _41
the number of _3112 is _3
the number of _3113 is _2
the number of _3115 is _1
the number of _3124 is _4
the number of _3133 is _1
the number of _3205 is _35
the number of _3207 is _1
the number of _3209 is _6
the number of _3300 is _1

the number of _3400 is _1
the number of _4111 is _1
the number of _4122 is _1
the number of _4124 is _5
the number of _4205 is _11
the number of _4206 is _3
the number of _4207 is _1
the number of _4300 is _4
the number of _5000 is _1
the number of _5110 is _1
the number of _5111 is _3
the number of _5124 is _2
the number of _5205 is _1
the number of _5206 is _2
the number of _5400 is _4
the number of _6111 is _3
the number of _6112 is _4
the number of _6209 is _1
the number of _7111 is _10
the number of _7113 is _2
the number of _7209 is _3
the number of _7400 is _1
the number of _8300 is _1
the number of _8400 is _2
the number of _8405 is _1
the number of _9000 is _1
the number of _9400 is _2
the number of _11124 is _2
the number of _11300 is _1
the number of _11400 is _1

the number of _12208 is_1
the number of _21110 is_1
the number of _31110 is_9

the number of _41110 is_1

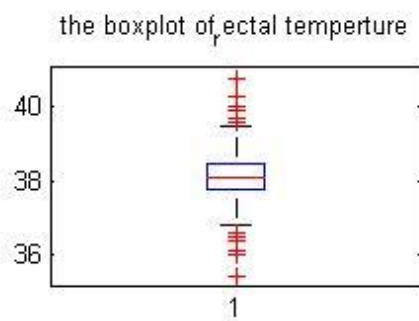
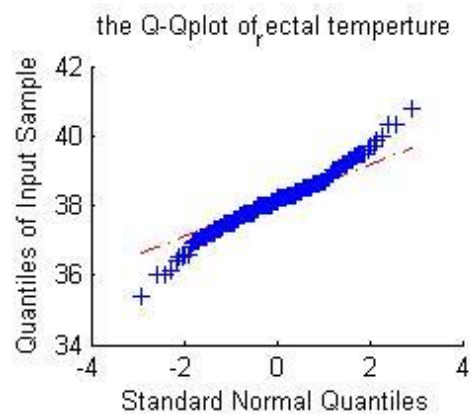
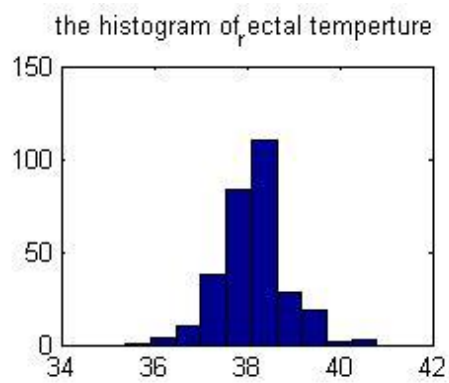
2) 7 组数值属性数据集的统计描述结果

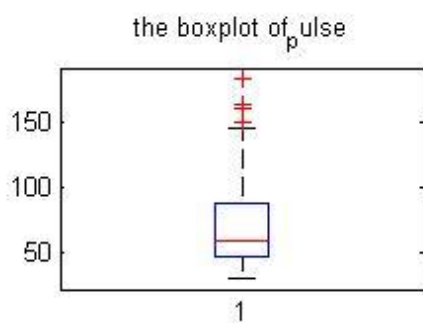
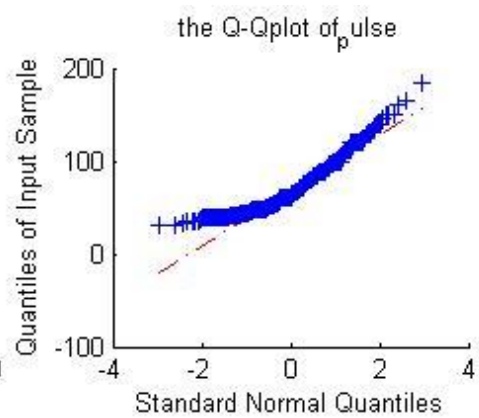
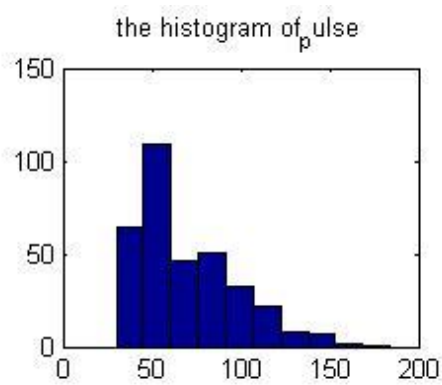
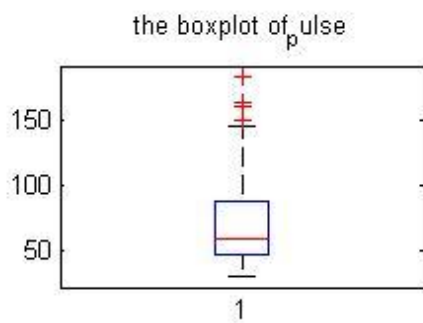
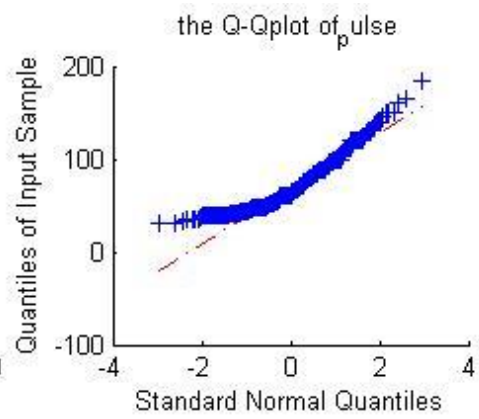
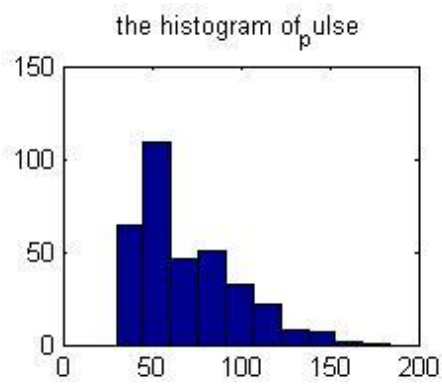
| 数据名称 | 描述属性 | 数值 | 数据名称 | 描述属性 | 数值 |
|--------------------------------|------|---------|--------------------------|------|---------|
| rectal temperture | 最大值 | 40.8 | pulse | 最大值 | 184 |
| | 最小值 | 35.4 | | 最小值 | 30 |
| | 均值 | 30.9842 | | 均值 | 65.7582 |
| | 中位数 | 38.1 | | 中位数 | 60 |
| | Q1 | 37.8 | | Q1 | 48 |
| | Q3 | 38.5 | | Q3 | 88 |
| | 缺失个数 | 69 | | 缺失个数 | 26 |
| respiratory | 最大值 | 96 | nasogastric reflux PH | 最大值 | 8.5 |
| | 最小值 | 8 | | 最小值 | 1 |
| | 均值 | 24.6332 | | 均值 | 0.93043 |
| | 中位数 | 28 | | 中位数 | 5.4 |
| | Q1 | 18 | | Q1 | 3.375 |
| | Q3 | 36 | | Q3 | 6.5 |
| | 缺失个数 | 71 | | 缺失个数 | 299 |
| packed cell volume | 最大值 | 75 | total protein | 最大值 | 89 |
| | 最小值 | 4 | | 最小值 | 3.3 |
| | 均值 | 41.0663 | | 均值 | 21.8766 |
| | 中位数 | 44 | | 中位数 | 7.5 |
| | Q1 | 37.125 | | Q1 | 6.5 |
| | Q3 | 52 | | Q3 | 58 |
| | 缺失个数 | 37 | | 缺失个数 | 43 |
| abdomcentesis total protein | 最大值 | 10.1 | | 最大值 | |
| | 最小值 | 0.1 | | 最小值 | |
| | 均值 | 1.0655 | | 均值 | |
| | 中位数 | 2.1 | | 中位数 | |
| | Q1 | 1.95 | | Q1 | |
| | Q3 | 3.9 | | Q3 | |
| | 缺失个数 | 235 | | 缺失个数 | |

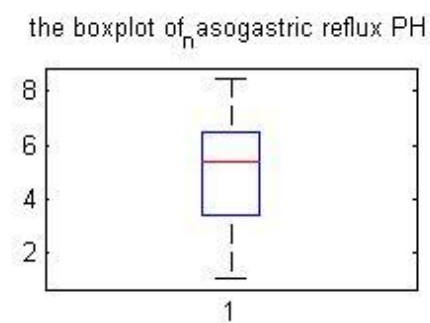
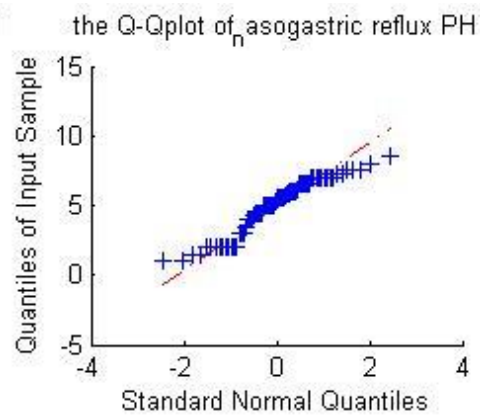
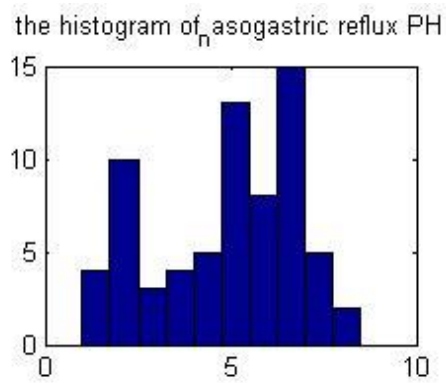
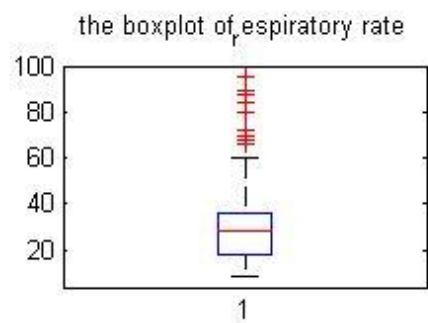
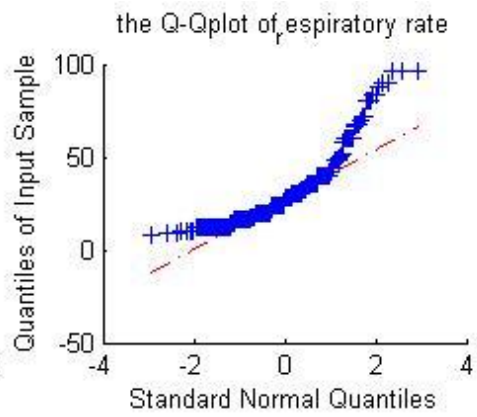
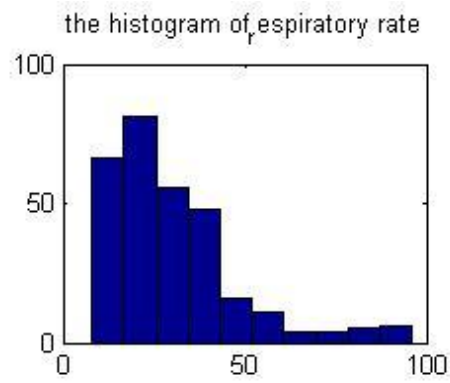
3) 7组数值属性数据集的结果图

数据组对应顺序如下:

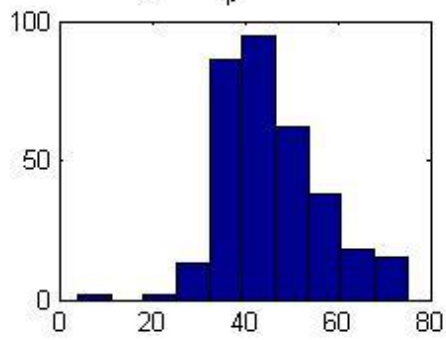
- 1.rectal temperature
- 2.pulse
- 3.respiratory rate
- 4.nasogastric reflux PH
- 5.packed cell volume
- 6.total protein
- 7.abdomocentesis total protein



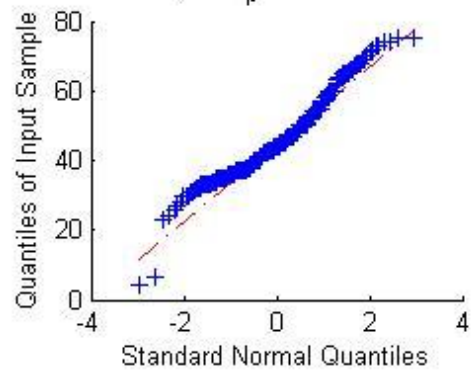




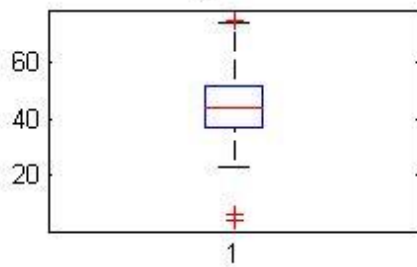
the histogram of p_{acked} cell volume



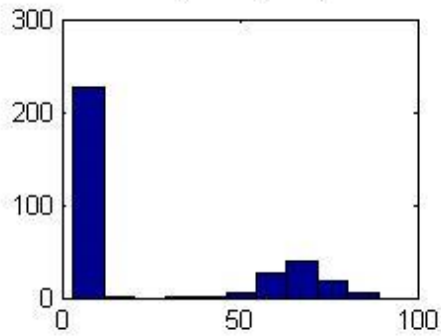
the Q-Qplot of p_{acked} cell volume



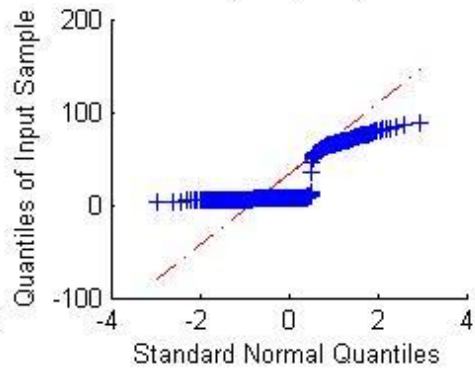
the boxplot of p_{acked} cell volume



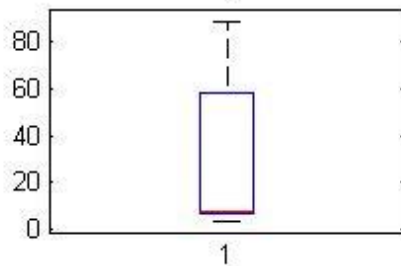
the histogram of t_{otal} protein



the Q-Qplot of t_{otal} protein



the boxplot of t_{otal} protein



四、 数据缺失值的处理

1. 将缺失部分剔除

通过 MATLAB 平台对于数据分析的过程来看，首先建立一个包含全部数据空间的矩阵，缺失值默认识别为 NAN，对每一个指标进行遍历，如果遇到缺失值则直接删去，位置由先一个数值进行填充，每个属性的最后空出来的部分补充为 NAN。最后输出为“re_delete.xlsx”文件。

代码保存为“re_delete.m”文件

核心代码如下：

```
data1(isnan(data1)==1)=-2;
for i=1:28
    for j=1:368
        del(i,j)=-1;
    end
end
for i=1:28
    k=1;
    for j=1:368
        if data1(i,j)>=-1
            del(i,k)=data1(i,j);
            k=k+1;
        end
    end
end
for i=1:28
    for j=1:368
        if del(i,j)==-1
            del(i,j)=nan;
        end
    end
end
xlswrite('re_delete.xlsx',del);
disp('re_delete.xls save succeed');
```

数据处理完之后，再次统计数值发现数据中间的缺失值直接被删除掉，数据量矩阵变小，可视化见下一部分。

2. 用最高频率值来填补缺失值

利用 MATLAB 中的 `tabulate` 函数，统计每一个属性中所有特征出现的次数，找出频率最高的相关数值。对原始数据读入的矩阵进行遍历，缺失值程序默认为 `NAN`，遍历过程中如果遇到 `NAN` 则根据其所属的属性及频率最高的特征进行替换。替换之后得到的矩阵保存为“`re_most.xlsx`”文件。

代码保存为“`re_most.m`”文件。

其中的核心代码如下图所示：

```
data1(isnan(data1)==1)=-1;
for i=1:28
    numdata=data(i,:);
    numtab=tabulate(numdata);
    [ma,1]=max(numtab(:,2));
    for j=1:368
        if data1(i,j)==-1
            data1(i,j)=numtab(1,1);
        end
    end
end
xlswrite('re_most.xlsx',data1);
disp('re_mosr save succeed!');
|
```

可视化见下一部分，通过缺失值补充可以看到数据有明显的频数改变。

3. 通过属性的相关关系来填补缺失值

通过对于属性的相关分析我通过计算各个属性之间的皮尔森相关系数，来确定两个属性之间的关联关系，计算公式如下：

$$\rho = \frac{\sum (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum (x_n - \bar{x})^2 \sum (y_n - \bar{y})^2}}$$

ρ 为皮尔森相关系数， \bar{x} 和 \bar{y} 分别为序列 x , y 的均值。

通过利用 MATLAB 的 corr 函数，计算得到 28*28 的矩阵，以 $p < 0.05$ 作为有显著性（对角线为自身的相关关系，值为 0，为了方便计算则重新赋值为 1），选择列向量里面的最小值（必须小于 0.05）所对应的的属性确定为目标属性的最相关属性。同时利用线性模拟函数寻找目标属性与最关联属性的拟合参数 K,B。得到如下的参数表格：

| Y（目标属性） | X（最关联属性） | K | B |
|---------|----------|----------|----------|
| 1 | 24 | 0.63641 | 0.536004 |
| 2 | 3 | 9.2E-07 | 0.585141 |
| 3 | 2 | 496466.7 | 313670 |
| 4 | 24 | 5.13002 | 23.95834 |
| 5 | 2 | 6.140284 | 55.8803 |
| 6 | 5 | 0.232976 | 9.329513 |
| 7 | 8 | 0.461636 | 1.343102 |
| 8 | 7 | 0.417177 | 0.783294 |
| 9 | 10 | 0.793319 | 1.619617 |
| 10 | 9 | 0.130304 | 0.864319 |
| 11 | 13 | 0.529632 | 1.533843 |
| 12 | 13 | 0.455952 | 1.742732 |
| 13 | 11 | 0.345126 | 1.058739 |
| 14 | 15 | 0.334349 | 0.896968 |
| 15 | 16 | 0.215363 | 1.001775 |
| 16 | 15 | 0.683497 | 0.475389 |
| 17 | 18 | 0.312891 | 1.223278 |
| 18 | 17 | 0.485138 | 1.561171 |
| 19 | 9 | 3.627882 | 32.59796 |
| 20 | 28 | 23.58141 | -17.3403 |
| 21 | 22 | 0.280981 | 0.84744 |
| 22 | 21 | 0.575672 | 0.812068 |
| 23 | 9 | 0.100683 | 1.284003 |
| 24 | 1 | 0.550411 | 0.597793 |
| 25 | 24 | -2788.97 | 7470.504 |
| 26 | 27 | 1.368148 | 88.76022 |
| 27 | 26 | 0.044221 | 1.714467 |
| 28 | 22 | -0.09727 | 1.70457 |

通过以下计算公式进行缺失值的补充：

$$Y(\text{目标属性}) = K * X(\text{最关联属性}) + B$$

核心代码如下：

皮尔森相关系数计算

```
for i=1:28
    val1=data1(i,:)' ;
    for j=1:28
        val2=data1(j,:)' ;
        [r(i,j),p(i,j)]=corr(val1,val2);
    end
end
for i=1:28
    p(i,i)=1;
end
disp('corr compute succeed');
```

寻找目标属性最相关属性的位置

```
loc=zeros(28,4);
for k=1:28
    vmin=min(p(k,:));
    lmin=find(p(k,:)==vmin);
    if vmin<0.05
        loc(k,1)=k;
        loc(k,2)=lmin;
    end
end
disp('corr_subject succeed');
```

线性拟合的参数计算

```
for k=1:28
    nh1=data1(loc(k,1),:);
    nh2=data2(loc(k,2),:);
    loc(k,3:4)=polyfit(nh2,nh1,1);
end
disp('polyfit succeed');
```

缺失值补充

```
for i=1:28
    k=loc(i,3);
    b=loc(i,4);
    x1=loc(i,2);
    for j=1:368
        if data2(i,j)==-1
            x=data1(i,x1);
            data(i,j)=k*x+b;
        end
    end
end
disp('re_corr succeed');
xlswrite('re_corr.xlsx',data);
```

通过缺失值填补之后，可视化见下一部分，通过比较可知，补缺前数据缺失值有 1927 个，缺失补充之后缺失值个数为 0。所以填补效果很明显。

补充之后得到的矩阵保存为“re_corr.xlsx”文件。

代码保存为“re_corr.m”文件。

4. 通过数据对象之间的相似性来填补缺失值

通过利用各数据对象之间的相似性来填补缺失值的过程中，先建立评分标准，在评分标准中以最小值表示相似程度最大，根据相似程度最大的对象的值进行目标值的填补。

评分标准如下：对于对标称属性值而言，如果相等则得分为 0，不等则得分为 1，但是如果都是缺失值则得分为 100；对于数值性数值而言，我们将两者进行做差后取绝对值作为得分。如果都为 NAN 则得分为 100。则这样相似度最大且无缺失的得分能够保证最小，所以以此来进行缺失值的补充。

补充之后得到的矩阵保存为 “re_subject.xlsx” 文件。

代码保存为 “re_subject.m” 文件。

其中核心代码如图所示：

```
clear all;
dat=xlsread('data_join.xlsx');
data=xlsread('data_join.xlsx');
data1=xlsread('data_join.xlsx');
data2=xlsread('data_join.xlsx');
listname={'surgery','age','hospital number','rectal temperture','pulse'...
'respiratory rate','temperature of extremities','peripheral pulse','mucous membranes','capillary refill time'...
'pain','peristalsis','abdominal distension','nasogastric tube','nasogastric reflux'...
'nasogastric reflux PH','rectal examination','abdomen','packed cell volume','total protein'...
'abdominocentesis appearance','abdomcentesis total protein','outcome','surgical lesion','type of lesion'...
'type of lesion 26','type of lesion 27','cp_data'};
numlist=[1,2,7,8,9,10,11,12,13,14,15,17,18,21,23,24,25,26,27,28];
vallist=[4,5,6,16,19,20,22];
num=20;
val=7;
data1(isnan(data1))==1=0;
data2(isnan(data2))==1=-1;
```

```

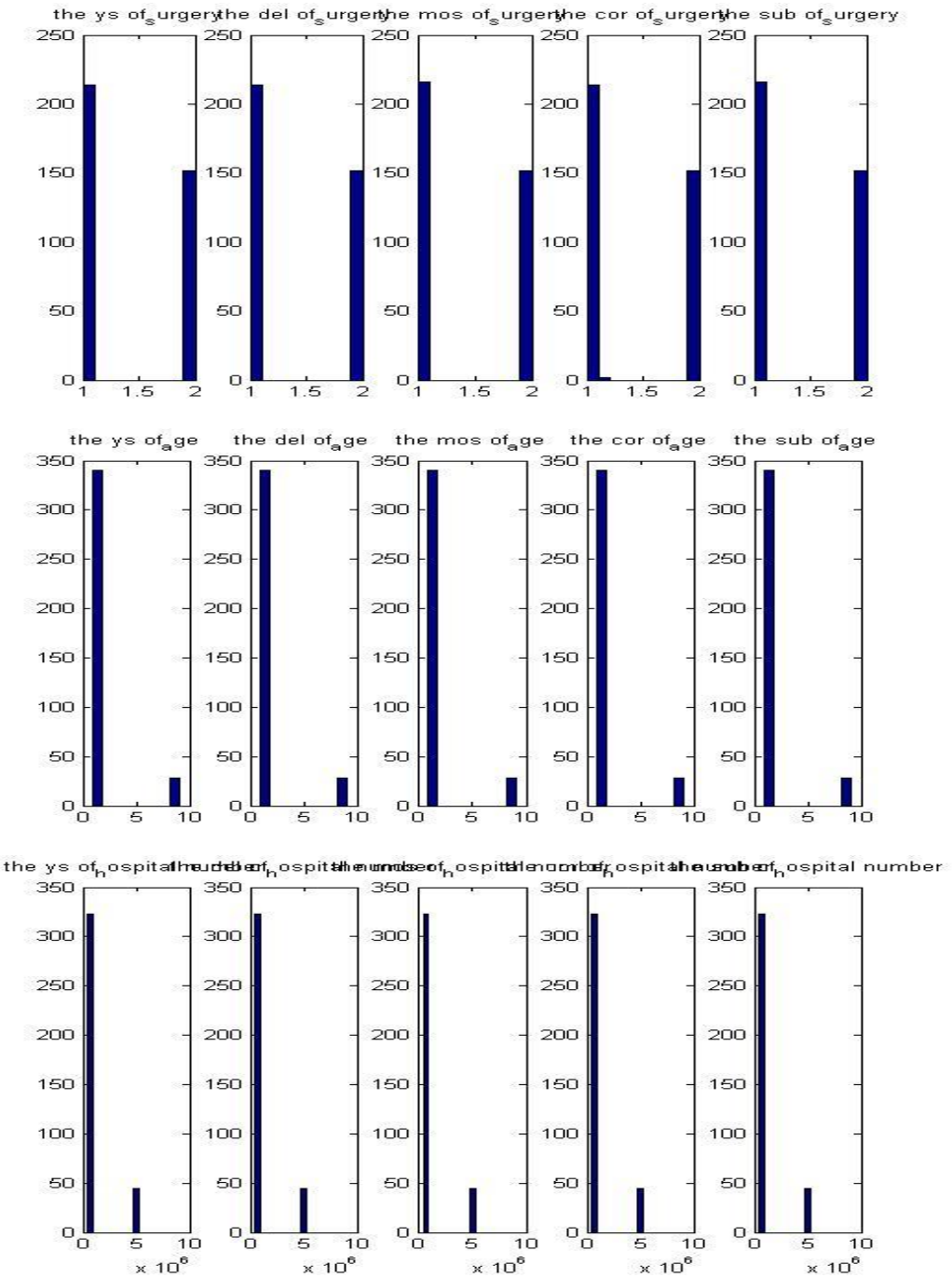
score=zeros(27,1);
score_sum=zeros(368,1);
for j=1:368
    sub2=data1(:,j);
    for n=1:20
        ncom1=sub1(numlist(n),1);
        ncom2=sub2(numlist(n),1);
        if ncom1==ncom2
            if ncom1==0
                score(n,1)=100;
            else
                score(n,1)=0;
            end
        else
            score(n,1)=1;
        end
    end
    for v=1:7
        vcom1=sub1(numlist(v),1);
        vcom2=sub2(numlist(v),1);
        if vcom1==vcom2
            if vcom1==0
                score(v+20,1)=100;
            else
                score(v+20,1)=abs(vcom1-vcom2);
            end
        else
            score(v+20,1)=abs(vcom1-vcom2);
        end
    end
    score_sum(j,1)=sum(score);
end
score_sum(i,1)=99999;
disp('score sum succeed!')
loc=find(score_sum==min(score_sum));
for k=1:28
    if data2(k,i)<0
        data(k,i)=data(k,min(loc));
    end
end
disp('data rewrite succeed!');
end

```

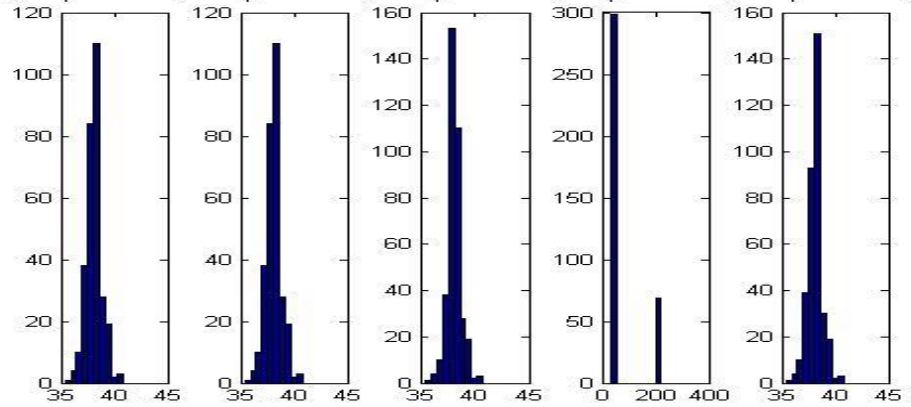
通过对比可以看出原先数据中的缺失值总共为 1927 个，经过补充之后的缺失值仅有 331 个，所以本次补充有一定的意义。

五、数据集缺失值补充可视化

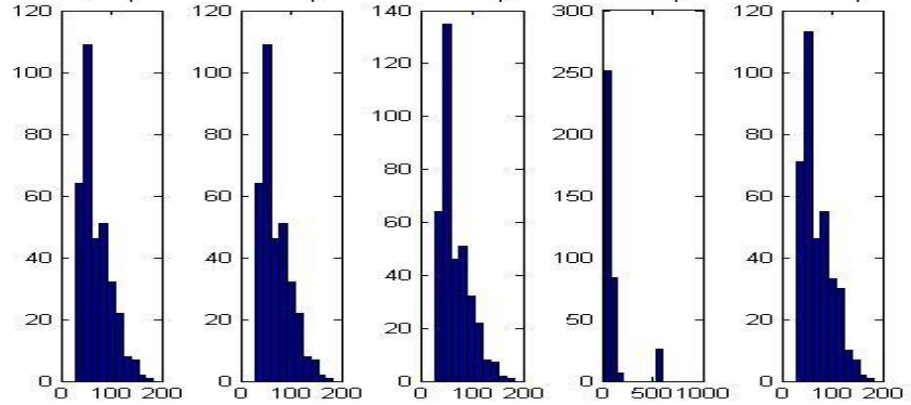
本数据一共 28 个属性，每个数据有 5 个直方统计图显示，ys 表示原始数据，del 表示删除缺失值，mos 表示以最大频数填充，cor 表示通过属性的相关关系来填补缺失值，sub 表示利用各数据对象之间的相似性来填补缺失值。



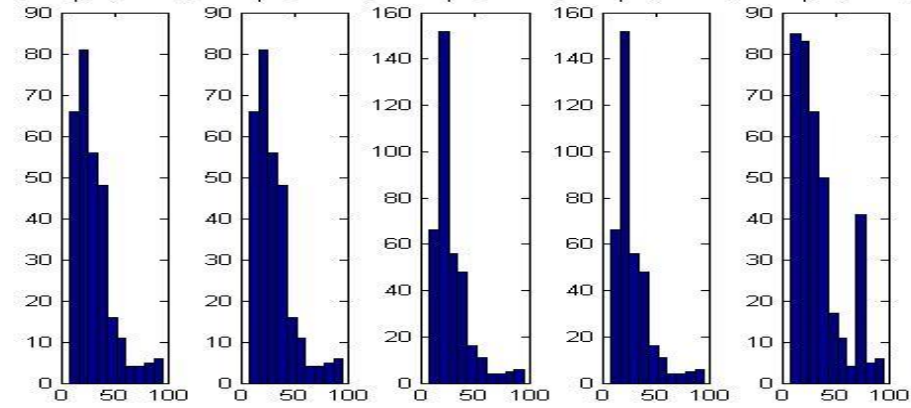
the ys of ectal temperature the del of ectal temperature the mos of ectal temperature the cor of ectal temperature the sub of ectal temperature



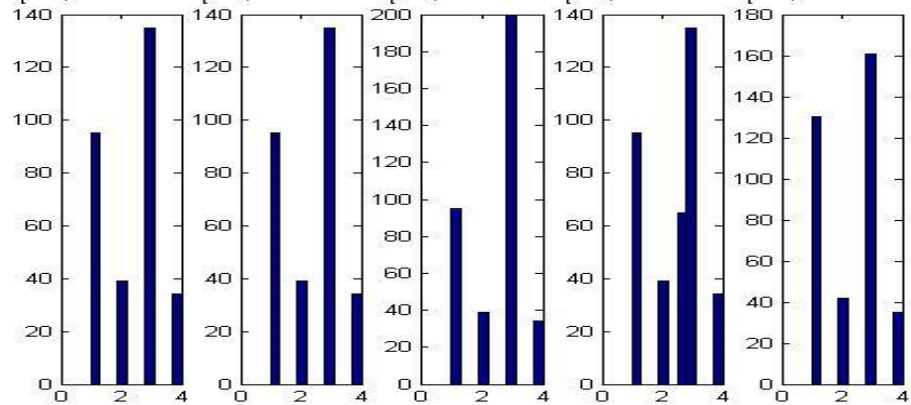
the ys of pulse the del of pulse the mos of pulse the cor of pulse the sub of pulse

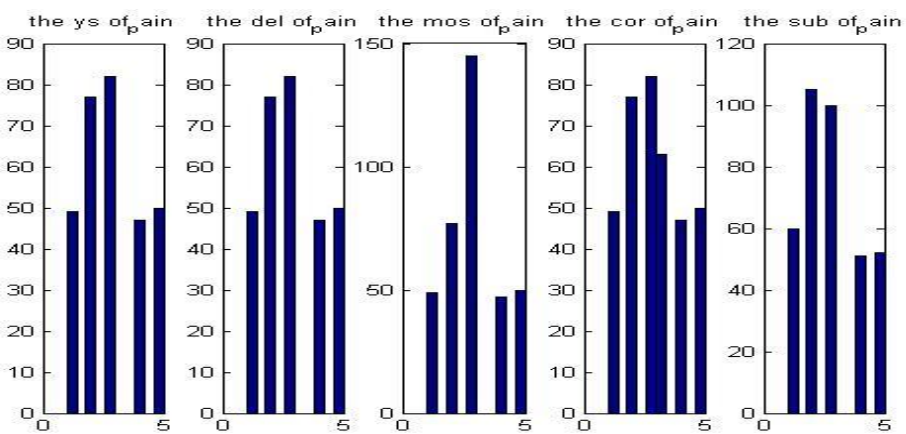
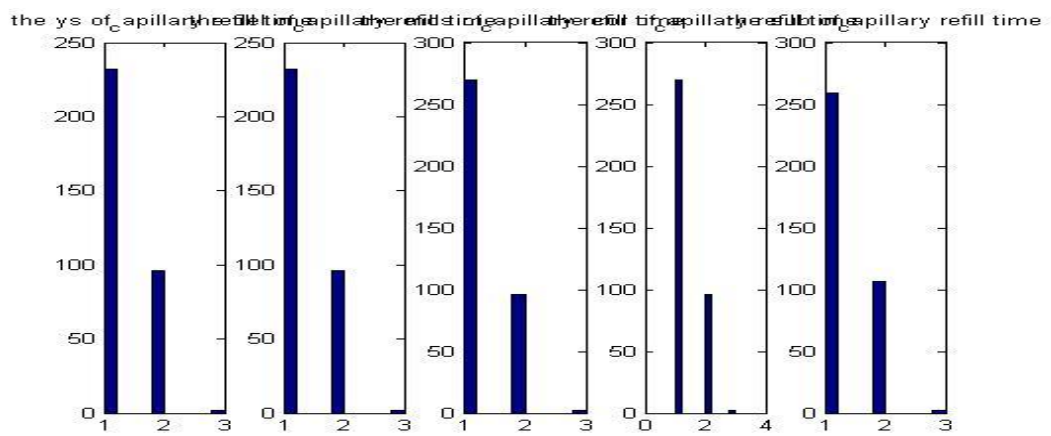
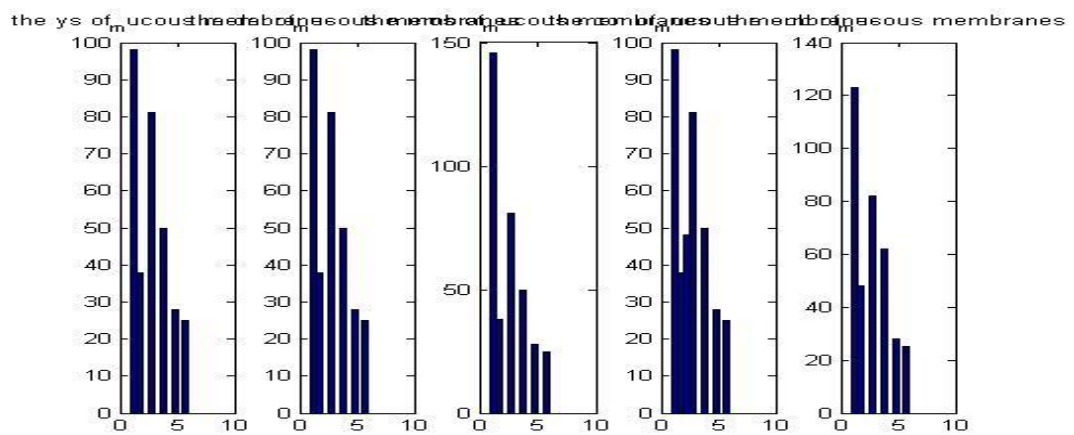
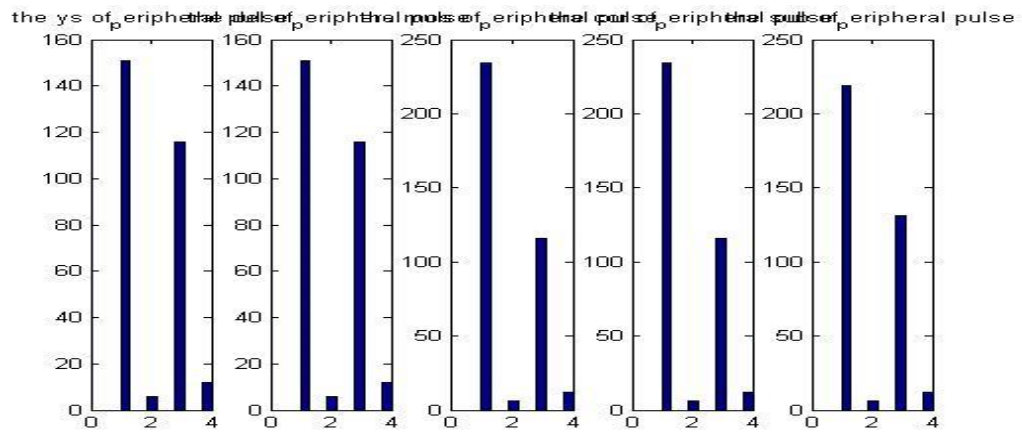


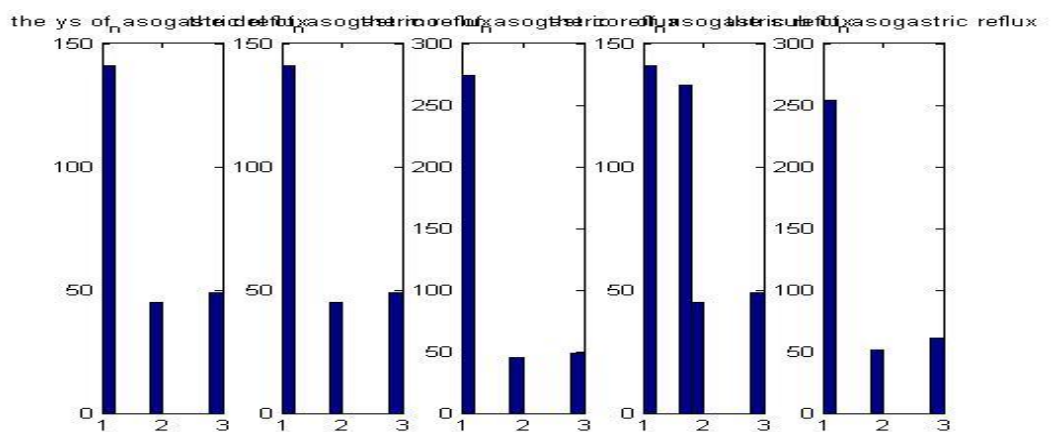
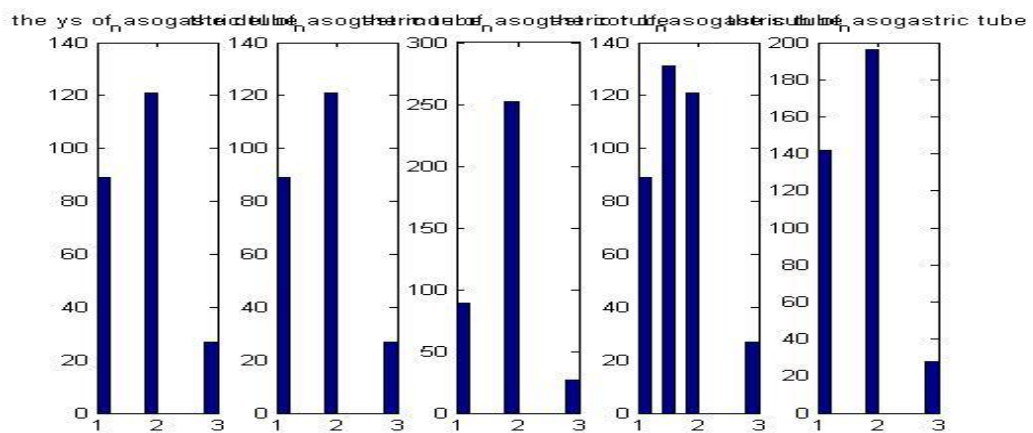
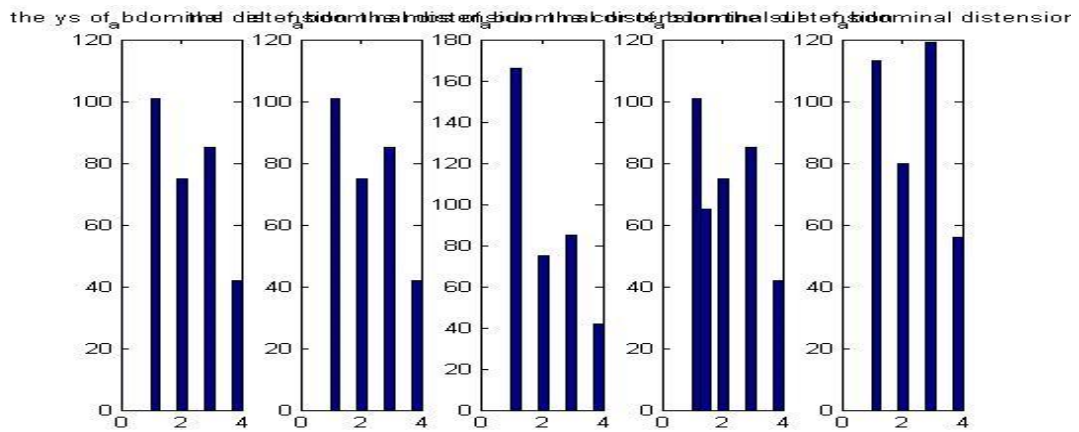
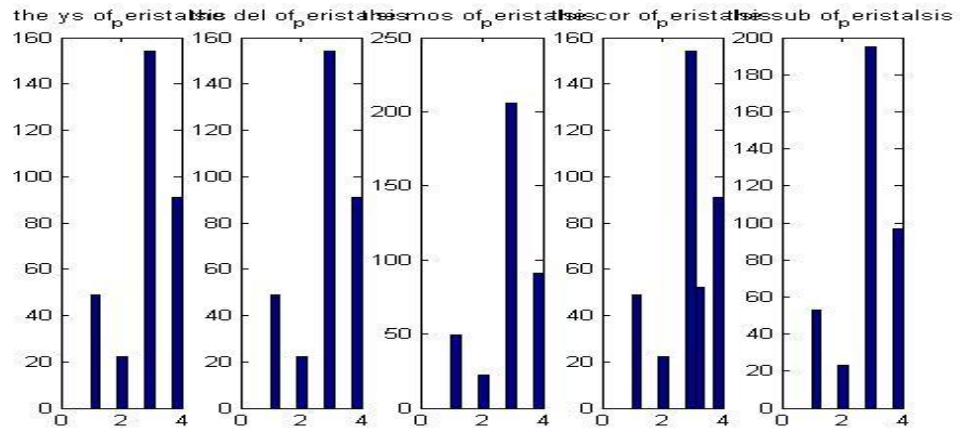
the ys of respiratory rate the del of respiratory rate the mos of respiratory rate the cor of respiratory rate the sub of respiratory rate

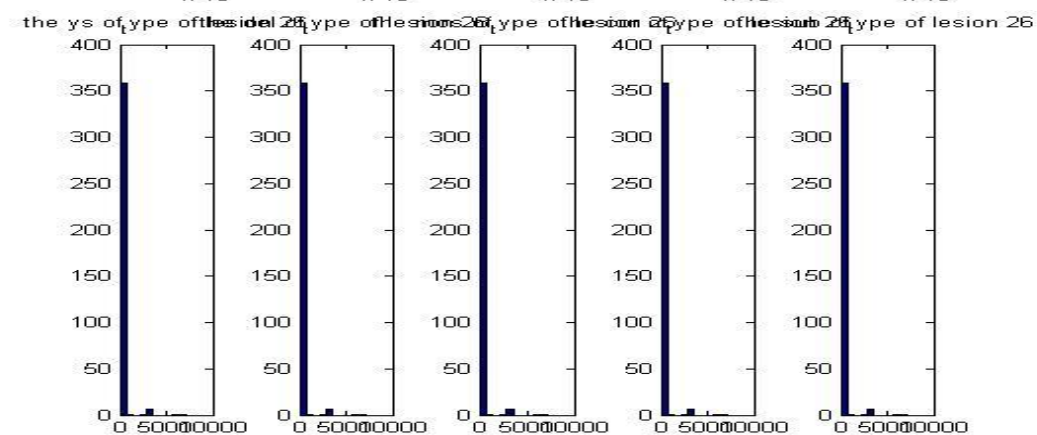
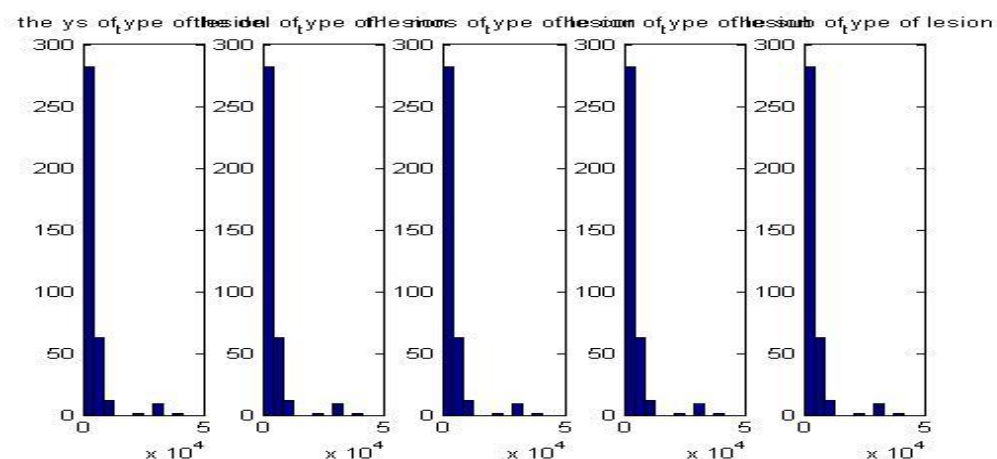
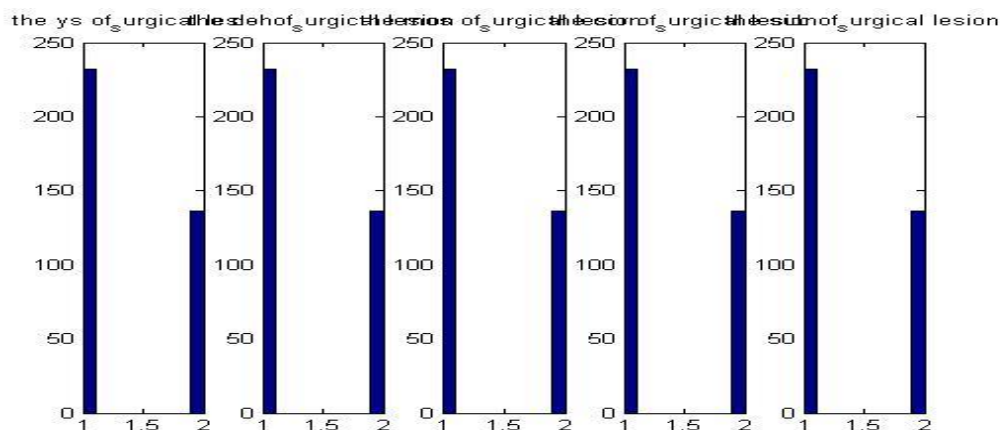


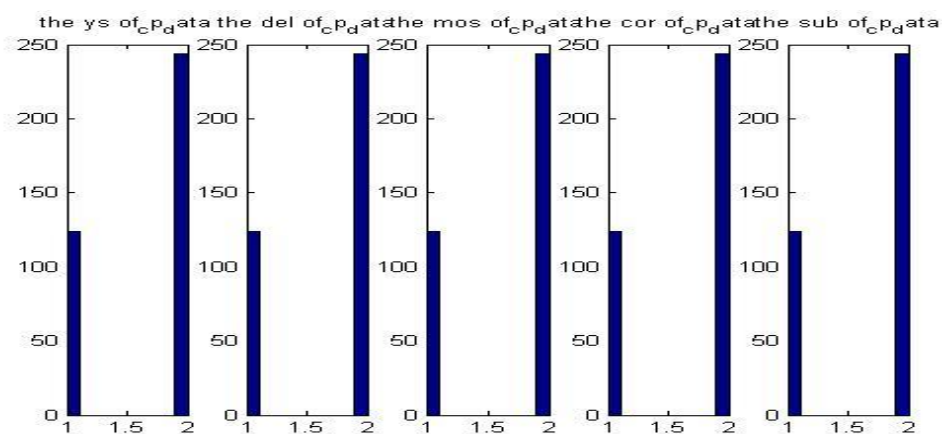
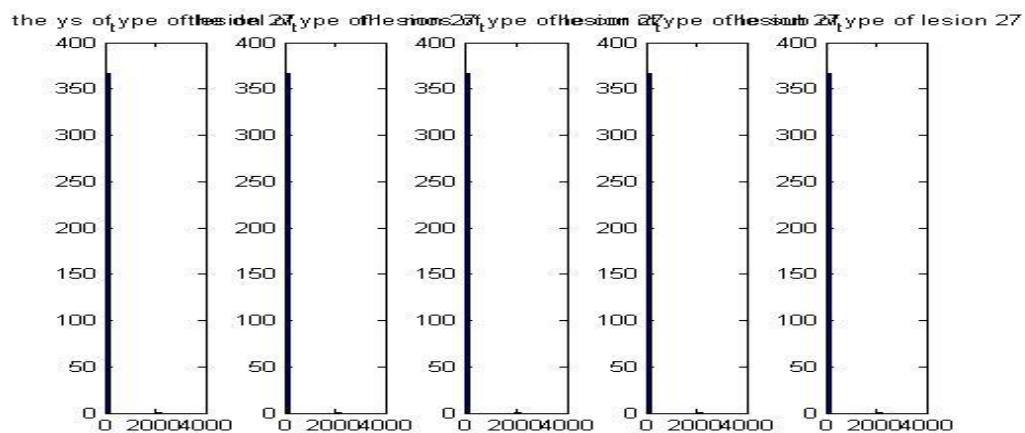
the ys of temperature of extremity the del of temperature of extremity the mos of temperature of extremity the cor of temperature of extremity the sub of temperature of extremity











六、 附录

原始数据结果输出原序列

1) 20 个对标称属性数据集的取值及对应频数的统计结果

| | |
|---------------------------------------|---------------------------------|
| <hr/> | <hr/> |
| 'in_surgery level' | 'in_pain level' |
| the number of_1 is_214 | the number of_1 is_49 |
| the number of_2 is_152 | the number of_2 is_77 |
| <hr/> | the number of_3 is_82 |
| <hr/> | the number of_4 is_47 |
| 'in_age level' | the number of_5 is_50 |
| the number of_1 is_340 | <hr/> |
| the number of_9 is_28 | <hr/> |
| <hr/> | 'in_peristalsis level' |
| <hr/> | the number of_1 is_49 |
| 'in_temperature of extremities level' | the number of_2 is_22 |
| the number of_1 is_95 | the number of_3 is_154 |
| the number of_2 is_39 | the number of_4 is_91 |
| the number of_3 is_135 | <hr/> |
| the number of_4 is_34 | <hr/> |
| <hr/> | 'in_abdominal distension level' |
| <hr/> | the number of_1 is_101 |
| 'in_peripheral pulse level' | the number of_2 is_75 |
| the number of_1 is_151 | the number of_3 is_85 |
| the number of_2 is_6 | the number of_4 is_42 |
| the number of_3 is_116 | <hr/> |
| the number of_4 is_12 | <hr/> |
| <hr/> | 'in_nasogastric tube level' |
| <hr/> | the number of_1 is_89 |
| 'in_mucous membranes level' | the number of_2 is_121 |
| the number of_1 is_98 | the number of_3 is_27 |
| the number of_2 is_38 | <hr/> |
| the number of_3 is_81 | <hr/> |
| the number of_4 is_50 | 'in_nasogastric reflux level' |
| the number of_5 is_28 | the number of_1 is_141 |
| the number of_6 is_25 | the number of_2 is_45 |
| <hr/> | the number of_3 is_49 |
| <hr/> | <hr/> |
| 'in_capillary refill time level' | <hr/> |
| the number of_1 is_232 | 'in_rectal examination level' |
| the number of_2 is_96 | the number of_1 is_68 |
| the number of_3 is_2 | the number of_2 is_14 |
| <hr/> | the number of_3 is_61 |

the number of _4 is_97

'in_abdomen level'

the number of _1 is_31

the number of _2 is_24

the number of _3 is_19

the number of _4 is_55

the number of _5 is_96

'in_abdominocentesis appearance level'

the number of _1 is_52

the number of _2 is_62

the number of _3 is_60

'in_outcome level'

the number of _1 is_225

the number of _2 is_89

the number of _3 is_52

'in_surgical lesion level'

the number of _1 is_232

the number of _2 is_136

'in_type of lesion level'

the number of _0 is_67

the number of _300 is_1

the number of _400 is_7

the number of _1111 is_1

the number of _1124 is_2

the number of _1400 is_10

the number of _2111 is_4

the number of _2112 is_6

the number of _2113 is_8

the number of _2124 is_9

the number of _2205 is_17

the number of _2206 is_5

the number of _2207 is_3

the number of _2208 is_23

the number of _2209 is_15

the number of _2300 is_2

the number of _2305 is_1

the number of _2322 is_2

the number of _3025 is_2

the number of _3111 is_41

the number of _3112 is_3

the number of _3113 is_2

the number of _3115 is_1

the number of _3124 is_4

the number of _3133 is_1

the number of _3205 is_35

the number of _3207 is_1

the number of _3209 is_6

the number of _3300 is_1

the number of _3400 is_1

the number of _4111 is_1

the number of _4122 is_1

the number of _4124 is_5

the number of _4205 is_11

the number of _4206 is_3

the number of _4207 is_1

the number of _4300 is_4

the number of _5000 is_1

the number of _5110 is_1

the number of _5111 is_3

the number of _5124 is_2

the number of _5205 is_1

the number of _5206 is_2

the number of _5400 is_4

the number of _6111 is_3

the number of _6112 is_4

the number of _6209 is_1

the number of _7111 is_10

the number of _7113 is_2

the number of _7209 is_3

the number of _7400 is_1

the number of _8300 is_1

the number of _8400 is_2

the number of _8405 is_1

the number of _9000 is_1

the number of _9400 is_2

the number of _11124 is_2

the number of _11300 is_1

| | |
|---|---|
| <hr/> <hr/> <hr/> <hr/> <hr/> | <hr/> <hr/> <hr/> <hr/> <hr/> |
| the number of _11400 is _1 the number of _12208 is _1 the number of _21110 is _1 the number of _31110 is _9 the number of _41110 is _1 'in_type of lesion 26 level' the number of _0 is _358 the number of _1400 is _1 the number of _2208 is _1 the number of _3111 is _3 the number of _3112 is _1 the number of _3205 is _2 | the number of _6112 is _1 the number of _7111 is _1 'in_type of lesion 27 level' the number of _0 is _367 the number of _2209 is _1 'in_cp_data level' the number of _1 is _124 the number of _2 is _244 |

2) 7 组数值属性数据集的统计描述结果

| | |
|---|--|
| <hr/> <hr/> <hr/> <hr/> <hr/> | <hr/> <hr/> <hr/> <hr/> <hr/> |
| 'the describe of _rectal temperture' the max value is _40.8 the min value is _35.4 the mean value is _30.9842 the Q2 value is _38.1 the Q1 value is _37.8 the Q3 value is _38.5 the NAN number is _69 | the max value is _96 the min value is _8 the mean value is _24.6332 the Q2 value is _28 the Q1 value is _18 the Q3 value is _36 the NAN number is _71 |
| <hr/> <hr/> <hr/> <hr/> <hr/> | <hr/> <hr/> <hr/> <hr/> <hr/> |
| disp the plot succeed! 'the describe of _pulse' the max value is _184 the min value is _30 the mean value is _65.7582 the Q2 value is _60 the Q1 value is _48 the Q3 value is _88 the NAN number is _26 | disp the plot succeed! 'the describe of _nasogastric reflux PH' the max value is _8.5 the min value is _1 the mean value is _0.93043 the Q2 value is _5.4 the Q1 value is _3.375 the Q3 value is _6.5 the NAN number is _299 |
| <hr/> <hr/> <hr/> <hr/> <hr/> | <hr/> <hr/> <hr/> <hr/> <hr/> |
| disp the plot succeed! 'the describe of _respiratory rate' | disp the plot succeed! 'the describe of _packed cell volume' the max value is _75 the min value is _4 the mean value is _41.0663 |

the Q2 value is_44
the Q1 value is_37.125
the Q3 value is_52
the NAN number is_37

disp the plot succeed!

'the describe of _total protein'

the max value is_89
the min value is_3.3
the mean value is_21.8766
the Q2 value is_7.5
the Q1 value is_6.5
the Q3 value is_58

the NAN number is_43

disp the plot succeed!

'the describe of _abdomcentesis total
protein'

the max value is_10.1
the min value is_0.1
the mean value is_1.0655
the Q2 value is_2.1
the Q1 value is_1.95
the Q3 value is_3.9
the NAN number is_235
