数据挖掘作业 1

数据探索性分析与预处理

（马的疝病分析）

姓名：雷丙震

学号：2120161300

日期：2016.4.17

目录

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. 数据分析要求

（1） 数据可视化和摘要

a) 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。

b) 数据的可视化（针对数值属性）

绘制直方图，如 mxPH，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

（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。

3. 算法实现

step1 读入数据

```
%% 读入数据
data=xlsread('horse-colic.xlsx');
%每列数据所对应的属性名称
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'};
%标称属性所在列
categorylist=[1,2,7,8,9,10,11,12,13,14,15,17,18,21,23,24,25,26,27,28];
category_num=size(categorylist,2);
%数值属性所在列
valuelist=[4,5,6,16,19,20,22];
%第 3 列 ID 号没有统计学意义，不在数据分析之内
value_num=size(valuelist,2);
```

```
[M,N]=size(data);
```

step2 数据摘要

在对数值属性进行计算统计数值时，对缺失值进行了置 0 处理

```
%% 数据摘要
disp('========================= 数 据 摘 要 ==========================');
%标称属性，给出每个可能取值的频数
disp('  对标称属性，给出每个可能取值的频数');
for i=1:category_num%标称属性的列数
    column=categorylist(i);%在原始数据中的列数
    numdata=data(:,column);
    numtab=tabulate(numdata);
    %统计矩阵中元素出现的次数，第一列为元素，第二列为次数，第三列为百分比
    disp('  ---------------------------------');
    [m,n]=size(numtab);
    disp(strcat(listname(column),'特征中:'));
    for j=1:m
        disp(strcat('  标 称 ',num2str(numtab(j,1)),' 的 频 数 是 ',num2str(numtab(j,2))));
    end
end
disp('************************ 我 是 分 割 线 **************************');
%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
disp('  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');
result2=zeros(column,7);
for i=1:value_num
    column=valuelist(i);
    valdata1=data(:,column);
    valdata=data(:,column);
    disp('  ----------------------------------------------------');
    disp(strcat('数值属性',listname(column),'的特征描述：'));
    result2(i,1)=max(valdata);%最大值
    disp(strcat('  最大值是',num2str(result2(i,1))));
    result2(i,2)=min(valdata);%最小值
    disp(strcat('  最小值是',num2str(result2(i,2))));
    valdata1(isnan(valdata1)==1)=0;%缺失值置 0
    result2(i,3)=mean(valdata1);%均值
```

```
        disp(strcat('        均值是',num2str(result2(i,3))));
        result2(i,4)=prctile(valdata,50);%中位数
        disp(strcat('        中位数是',num2str(result2(i,4))));
        result2(i,5)=prctile(valdata,25);%第一个四分位数
        disp(strcat('        Q1 值是',num2str(result2(i,5))));
        result2(i,6)=prctile(valdata,75);%第 3 四分位数
        disp(strcat('        Q3 值是',num2str(result2(i,6))));
        result2(i,7)=sum(isnan(valdata));%缺失值个数
        disp(strcat('        缺失值个数为',num2str(result2(i,7))));
    end
```

step3 数据可视化

```
    %% 数据可视化
    for i=1:value_num
        column=valuelist(i);
        valdata1=data(:,column);
        valdata=data(:,column);
        figure;subplot(221);hist(valdata);
        title(strcat(listname(column),'的直方图'));
        subplot(222);qqplot(valdata);
        title(strcat(listname(column),'的 QQ 图'));
        subplot(223);boxplot(valdata);
        title(strcat(listname(column),'的盒图'));
    end
```

step4 剔除缺失部分

```
    %% 将缺失值剔除
    [row,column]=find(isnan(data)==1);%缺失值所在位置坐标
    t=1;
    for i=1:M
        if ismember(i,row)==0
            data_delete(t,:)=data(i,:);
            t=t+1;
        end
    end
    xlswrite('data_deletemissing.xlsx',data_delete);
```

step5 用最高频率值填补缺失值

```
    %% 用最高频率值来填补缺失值
    data_mode=data;%对缺失数据置为-1
```

```
for column=1:N
    numdata=data(:,column);
    numtab=tabulate(numdata);
    %统计矩阵中元素出现的次数，第一列为元素，第二列为次数，第三列为百分比
    [ma,pos]=max(numtab(:,2));
    for row=1:M
        if isnan(data(row,column))==1
            data_mode(row,column)=numtab(pos,1);
        end
    end
end
xlswrite('data_mode.xlsx',data_mode);
```

step6 通过属性的相关关系来填补缺失值

计算各属性之间的皮尔逊相关系数，筛选出与待填补属性最相关的属性，然后采用一元线性回归分析，计算两者之间的拟合方程，由此方程计算缺失值。

```
%% 通过属性的相关关系来填补缺失值
corr_test=data;corr_test(:,3)=[];%剔除 ID 列
data_coor=data;data_coor(:,3)=[];%剔除 ID 列
r=zeros(27);p=zeros(27);
corr_test(isnan(corr_test)==1)=0;%NAN 数据置 0 以便计算相关性
for i=1:27
    for j=1:27
        feature1=corr_test(:,i);
        feature2=corr_test(:,j);
        [r(i,j),p(i,j)]=corr(feature1,feature2);%r 为相关系数,p 为置信区间
    end
end
for i=1:27
    p(i,i)=1;
end
sign=zeros(27,4);
%sign 用来存储相关和拟合的结果，第一列为 Y，第二列为 X，第三列为 k，第四列为 b
for k=1:27
    [pmin,qmin]=min(abs(p(k,:)));%置信区间最小的值
    sign(k,1)=k;%Y 所在的列
    sign(k,2)=qmin;%X 所在的列
end
for k=1:27
    vector1=corr_test(:,sign(k,1));%Y
    vector2=corr_test(:,sign(k,2));%X
```

```
        sign(k,3:4)=polyfit(vector2,vector1,1);%线性拟合系数
    end
    for column=1:27
        k=sign(column,3);
        b=sign(column,4);
        x=sign(column,2);
        for row=1:M
            if isnan(data_coor(row,column))==1
                data_coor(row,column)=k*data_coor(row,x)+b;
            end
        end
    end
    data_corr(:,1:2)=data_coor(:,1:2);
    data_corr(:,3)=data(:,3);
    data_corr(:,4:28)=data_coor(:,3:27);
    xlswrite('data_corr.xlsx',data_corr);
```

step7 通过数据对象之间的相似性来填补缺失值

计算各属性之间的余弦相似性，筛选出与待填补属性最相似的属性，然后用此属性的对应值填补缺失数据。

```
    %% 通过数据对象之间的相似性来填补缺失值
    cos_sim=data;cos_sim(:,3)=[];%剔除 ID 列
    data_cos=data;data_cos(:,3)=[];%剔除 ID 列
    sim=zeros(27);
    cos_sim(isnan(cos_sim)==1)=0;%NAN 数据置 0 以便计算相关性
    %使用余弦相似性进行相似性度量
    for i=1:27
        for j=1:27
            feature1=cos_sim(:,i);
            feature2=cos_sim(:,j);
            sim(i,j)=sum(feature1.*feature2)/...

(sqrt(sum(feature1.*feature1))*sqrt(sum(feature2.*feature2)));
        end
    end
    for i=1:27
        sim(i,i)=0;
    end
    sign_sim=zeros(27,2);
    %sign 用来存储相关和拟合的结果，第一列为 Y，第二列为 X
    for k=1:27
        [smax,pmax]=max(sim(k,:));%余弦相似性最大值
        sign_sim(k,1)=k;%Y 所在的列
```

```
        sign_sim(k,2)=pmax;%X 所在的列
    end
    for column=1:27
        x=sign_sim(column,2);
        for row=1:M
            if isnan(data_cos(row,column))==1
                data_cos(row,column)=data_cos(row,x);
            end
        end
    end
    data_sim(:,1:2)=data_cos(:,1:2);
    data_sim(:,3)=data(:,3);
    data_sim(:,4:28)=data_cos(:,3:27);
    xlswrite('data_sim.xlsx',data_sim);
```

step8 可视化比较新旧数据集

对标称属性，通过绘制新旧数据的散点图进行可视化；对数值属性，绘制新旧数据集均值的直方图进行可视化。

```
    %% 可视化比较新旧数据集
    for column=1:N
        %对标称属性绘制散点图进行比较
        if ismember(column,categorylist)==1
            %
            figure;subplot(221);plot(data(:,column),'r*');hold on;
            plot(data_delete(:,column),'b.');hold off;
            title(strcat(listname(column),'的散点图'));
            legend('pre_data','data_deletemissing');
            %
            subplot(222);plot(data(:,column),'r*');hold on;
            plot(data_mode(:,column),'b.');hold off;
            title(strcat(listname(column),'的散点图'));
            legend('pre_data','data_mode');
            %
            subplot(223);plot(data(:,column),'r*');hold on;
            plot(data_corr(:,column),'b.');hold off;
            title(strcat(listname(column),'的散点图'));
            legend('pre_data','data_corr');
            %
            subplot(224);plot(data(:,column),'r*');hold on;
            plot(data_sim(:,column),'b.');hold off;
            title(strcat(listname(column),'的散点图'));
            legend('pre_data','data_sim');
        elseif ismember(column,valuelist)==1
```

```
                %对数值属性比较均值
                %原始数据
                valdata1=data(:,column);
                valdata1(isnan(valdata1)==1)=0;%缺失值置 0
                pre_mean=mean(valdata1);%均值
                %剔除缺失值
                valdata1=data_delete(:,column);
                valdata1(isnan(valdata1)==1)=0;%缺失值置 0
                delete_mean=mean(valdata1);%均值
                %众数补齐
                valdata1=data_mode(:,column);
                valdata1(isnan(valdata1)==1)=0;%缺失值置 0
                mode_mean=mean(valdata1);%均值
                %相关补齐
                valdata1=data_corr(:,column);
                valdata1(isnan(valdata1)==1)=0;%缺失值置 0
                corr_mean=mean(valdata1);%均值
                %相似性补齐
                valdata1=data_sim(:,column);
                valdata1(isnan(valdata1)==1)=0;%缺失值置 0
                sim_mean=mean(valdata1);%均值

    mean_data=[pre_mean,delete_mean,mode_mean,corr_mean,sim_mean];
                figure;bar(mean_data);
                set(gca,'XTickLabel',{'pre_mean','delete_mean','mode_mean',...
                    'corr_mean','sim_mean'});
            end
        end
```

4. 结果与讨论

（1） 数据摘要（step2 输出结果）

```
            ========================        数        据        摘        要
        ==========================
                对标称属性，给出每个可能取值的频数
                ----------------------------------
                  'surgery 特征中:'

                  标称 1 的频数是 214
                  标称 2 的频数是 152
                ----------------------------------
                  'age 特征中:'

                  标称 1 的频数是 340
                  标称 2 的频数是 0
```

标称 3 的频数是 0
标称 4 的频数是 0
标称 5 的频数是 0
标称 6 的频数是 0
标称 7 的频数是 0
标称 8 的频数是 0
标称 9 的频数是 28

'temperature of extremities 特征中:'

标称 1 的频数是 95
标称 2 的频数是 39
标称 3 的频数是 135
标称 4 的频数是 34

'peripheral pulse 特征中:'

标称 1 的频数是 151
标称 2 的频数是 6
标称 3 的频数是 116
标称 4 的频数是 12

'mucous membranes 特征中:'

标称 1 的频数是 98
标称 2 的频数是 38
标称 3 的频数是 81
标称 4 的频数是 50
标称 5 的频数是 28
标称 6 的频数是 25

'capillary refill time 特征中:'

标称 1 的频数是 232
标称 2 的频数是 96
标称 3 的频数是 2

'pain 特征中:'

标称 1 的频数是 49
标称 2 的频数是 77
标称 3 的频数是 82
标称 4 的频数是 47
标称 5 的频数是 50

'peristalsis 特征中:'

标称 1 的频数是 49
标称 2 的频数是 22
标称 3 的频数是 154
标称 4 的频数是 91

'abdominal distension 特征中:'

标称 1 的频数是 101
标称 2 的频数是 75
标称 3 的频数是 85
标称 4 的频数是 42

'nasogastric tube 特征中:'

标称 1 的频数是 89
标称 2 的频数是 121
标称 3 的频数是 27

'nasogastric reflux 特征中:'

标称 1 的频数是 141
标称 2 的频数是 45
标称 3 的频数是 49

'rectal examination 特征中:'

标称 1 的频数是 68
标称 2 的频数是 14
标称 3 的频数是 61
标称 4 的频数是 97

'abdomen 特征中:'

标称 1 的频数是 31
标称 2 的频数是 24
标称 3 的频数是 19
标称 4 的频数是 55
标称 5 的频数是 96

'abdominocentesis appearance 特征中:'

标称 1 的频数是 52
标称 2 的频数是 62
标称 3 的频数是 60

'outcome 特征中:'

标称 1 的频数是 225
标称 2 的频数是 89
标称 3 的频数是 52

'surgical lesion 特征中:'

标称 1 的频数是 232
标称 2 的频数是 136

'type of lesion 特征中:'

标称 0 的频数是 67
标称 300 的频数是 1
标称 400 的频数是 7
标称 1111 的频数是 1
标称 1124 的频数是 2
标称 1400 的频数是 10
标称 2111 的频数是 4
标称 2112 的频数是 6
标称 2113 的频数是 8
标称 2124 的频数是 9
标称 2205 的频数是 17
标称 2206 的频数是 5
标称 2207 的频数是 3
标称 2208 的频数是 23
标称 2209 的频数是 15
标称 2300 的频数是 2
标称 2305 的频数是 1
标称 2322 的频数是 2
标称 3025 的频数是 2
标称 3111 的频数是 41
标称 3112 的频数是 3
标称 3113 的频数是 2
标称 3115 的频数是 1
标称 3124 的频数是 4
标称 3133 的频数是 1
标称 3205 的频数是 35
标称 3207 的频数是 1

标称 3209 的频数是 6
标称 3300 的频数是 1
标称 3400 的频数是 1
标称 4111 的频数是 1
标称 4122 的频数是 1
标称 4124 的频数是 5
标称 4205 的频数是 11
标称 4206 的频数是 3
标称 4207 的频数是 1
标称 4300 的频数是 4
标称 5000 的频数是 1
标称 5110 的频数是 1
标称 5111 的频数是 3
标称 5124 的频数是 2
标称 5205 的频数是 1
标称 5206 的频数是 2
标称 5400 的频数是 4
标称 6111 的频数是 3
标称 6112 的频数是 4
标称 6209 的频数是 1
标称 7111 的频数是 10
标称 7113 的频数是 2
标称 7209 的频数是 3
标称 7400 的频数是 1
标称 8300 的频数是 1
标称 8400 的频数是 2
标称 8405 的频数是 1
标称 9000 的频数是 1
标称 9400 的频数是 2
标称 11124 的频数是 2
标称 11300 的频数是 1
标称 11400 的频数是 1
标称 12208 的频数是 1
标称 21110 的频数是 1
标称 31110 的频数是 9
标称 41110 的频数是 1

'type of lesion 26 特征中:'

标称 0 的频数是 358
标称 1400 的频数是 1
标称 2208 的频数是 1
标称 3111 的频数是 3
标称 3112 的频数是 1

标称 3205 的频数是 2
标称 6112 的频数是 1
标称 7111 的频数是 1

'type of lesion 27 特征中:'

标称 0 的频数是 367
标称 2209 的频数是 1

'cp_data 特征中:'

标称 1 的频数是 124
标称 2 的频数是 244

********** 我 是 分 割 线 **********

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

'数值属性 rectal temperture 的特征描述：'

最大值是 40.8
最小值是 35.4
均值是 30.9842
中位数是 38.1
Q1 值是 37.8
Q3 值是 38.5
缺失值个数为 69

'数值属性 pulse 的特征描述：'

最大值是 184
最小值是 30
均值是 65.7582
中位数是 60
Q1 值是 48
Q3 值是 88
缺失值个数为 26

'数值属性 respiratory rate 的特征描述：'

最大值是 96
最小值是 8
均值是 24.6332
中位数是 28

Q1 值是 18
Q3 值是 36
缺失值个数为 71

'数值属性 nasogastric reflux PH 的特征描述：'

最大值是 8.5
最小值是 1
均值是 0.93043
中位数是 5.4
Q1 值是 3.375
Q3 值是 6.5
缺失值个数为 299

'数值属性 packed cell volume 的特征描述：'

最大值是 75
最小值是 4
均值是 41.0663
中位数是 44
Q1 值是 37.125
Q3 值是 52
缺失值个数为 37

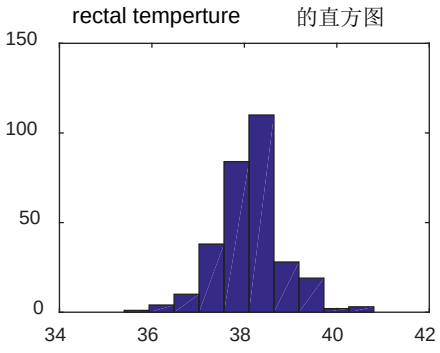
'数值属性 total protein 的特征描述：'

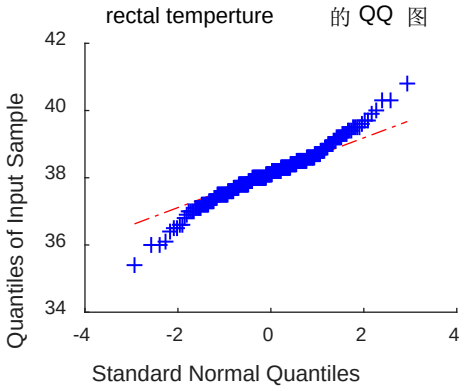
最大值是 89
最小值是 3.3
均值是 21.8766
中位数是 7.5
Q1 值是 6.5
Q3 值是 58
缺失值个数为 43

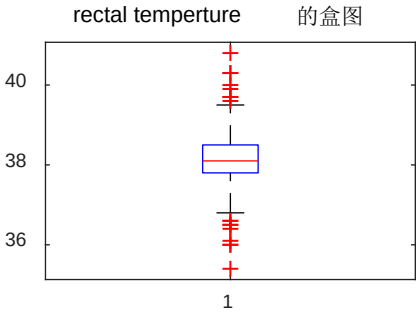
'数值属性 abdomcentesis total protein 的特征描述：'

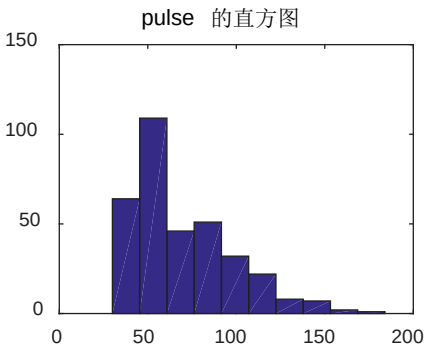
最大值是 10.1
最小值是 0.1
均值是 1.0655
中位数是 2.1
Q1 值是 1.95
Q3 值是 3.9
缺失值个数为 235

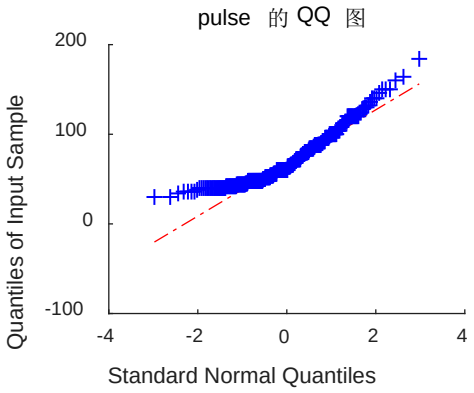
（2） 数据可视化（step3 输出结果）

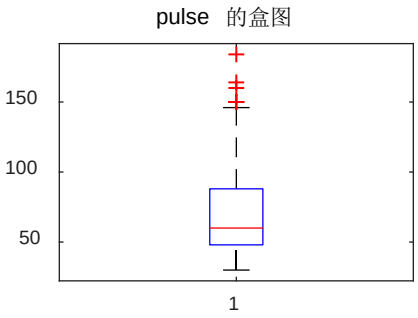

rectal temperture 的直方图
150
100
50
0
34
36
38
40
42

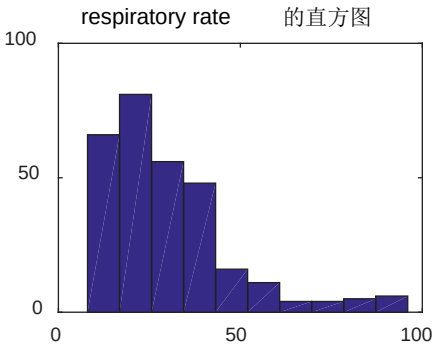

rectal temperture 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
42
40
38
36
34
-4
-2
0
2
4

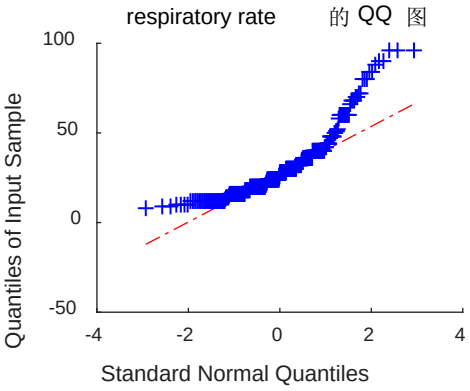

rectal temperture 的盒图
40
38
36
1

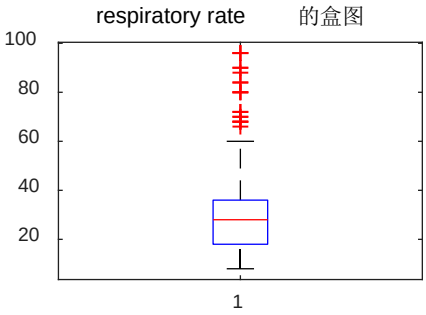

pulse 的直方图
150
100
50
0
0
50
100
150
200

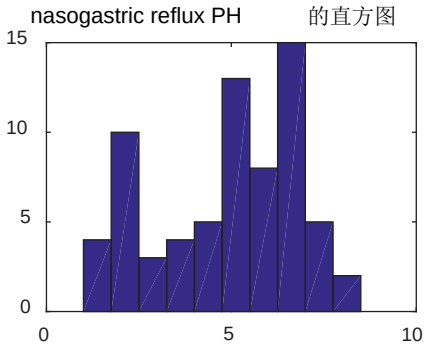

pulse 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
200
100
0
-100
-4
-2
0
2
4

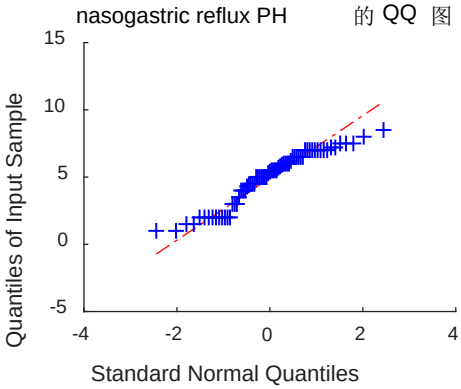

pulse 的盒图
150
100
50
1

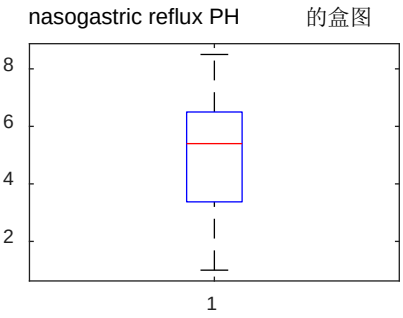

respiratory rate 的直方图
100
50
0
0
50
100

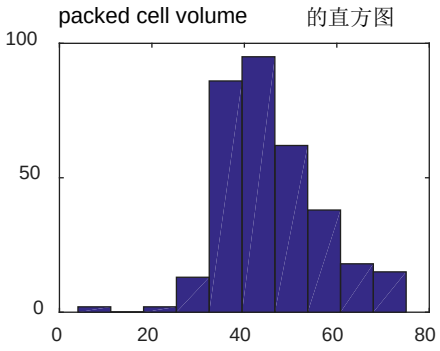

respiratory rate 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
100
50
0
-50
-4
-2
0
2
4

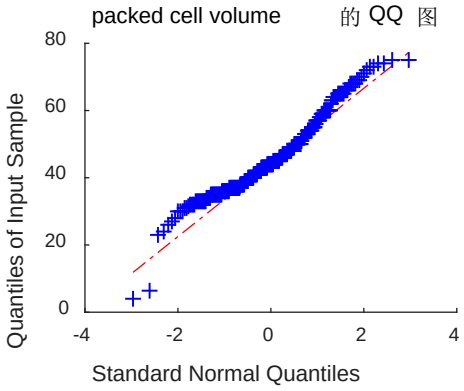

respiratory rate 的盒图
100
80
60
40
20
1

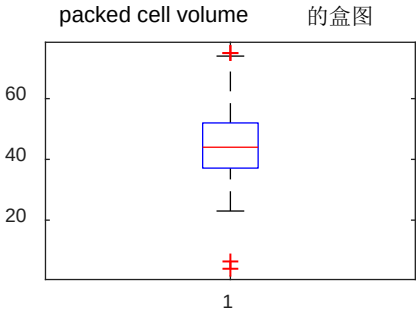

nasogastric reflux PH 的直方图
15
10
5
0
0
5
10

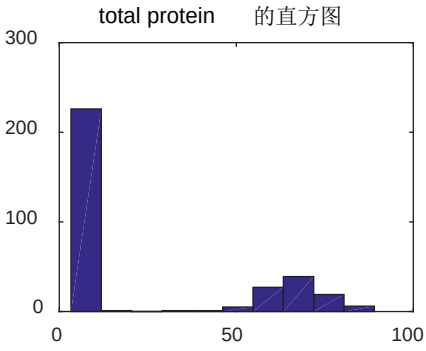

nasogastric reflux PH 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
15
10
5
0
-5
-4
-2
0
2
4

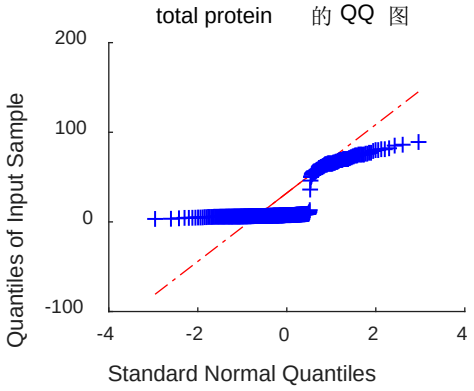

nasogastric reflux PH 的盒图
8
6
4
2
1

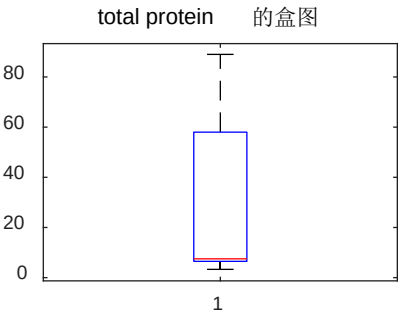

packed cell volume 的直方图
100
50
0
0
20
40
60
80

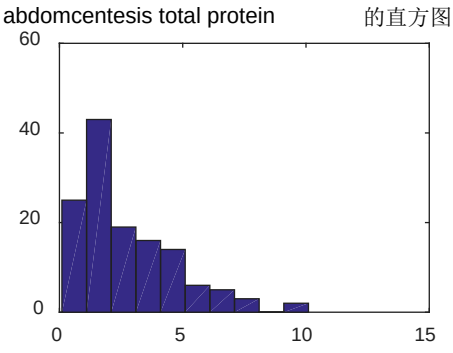

packed cell volume 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
80
60
40
20
0
-4
-2
0
2
4

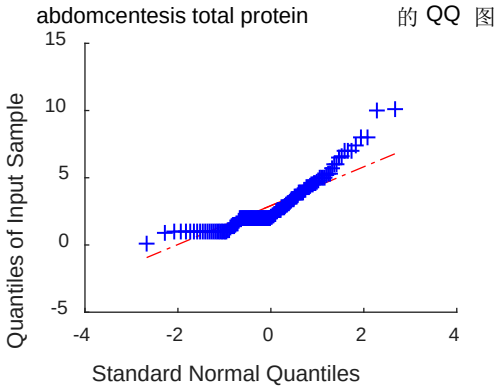

packed cell volume 的盒图
60
40
20
1

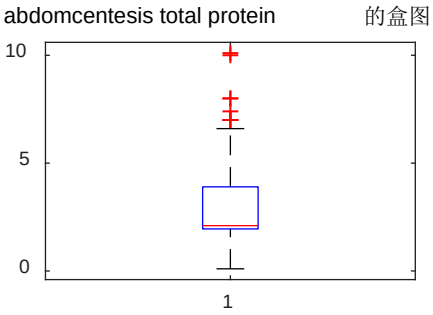

total protein 的直方图
300
200
100
0
0
50
100


total protein 的 QQ 图
Quantiles of Input Sample
Standard Normal Quantiles
200
100
0
-100
-4
-2
0
2
4


total protein 的盒图
80
60
40
20
0
1







讨论：由直方图和 QQ 图可以看出，在对马的疝病分析的 27 个特征的 7 个数值特征中，只有'rectal temperture'、'nasogastric reflux PH'和'packed cell volume'三个数值属性大致呈正态分布。

（3） 数据缺失值处理（step4~step7 输出结果）

原始数据缺失数目：1297

删除缺失值后缺失数据：0

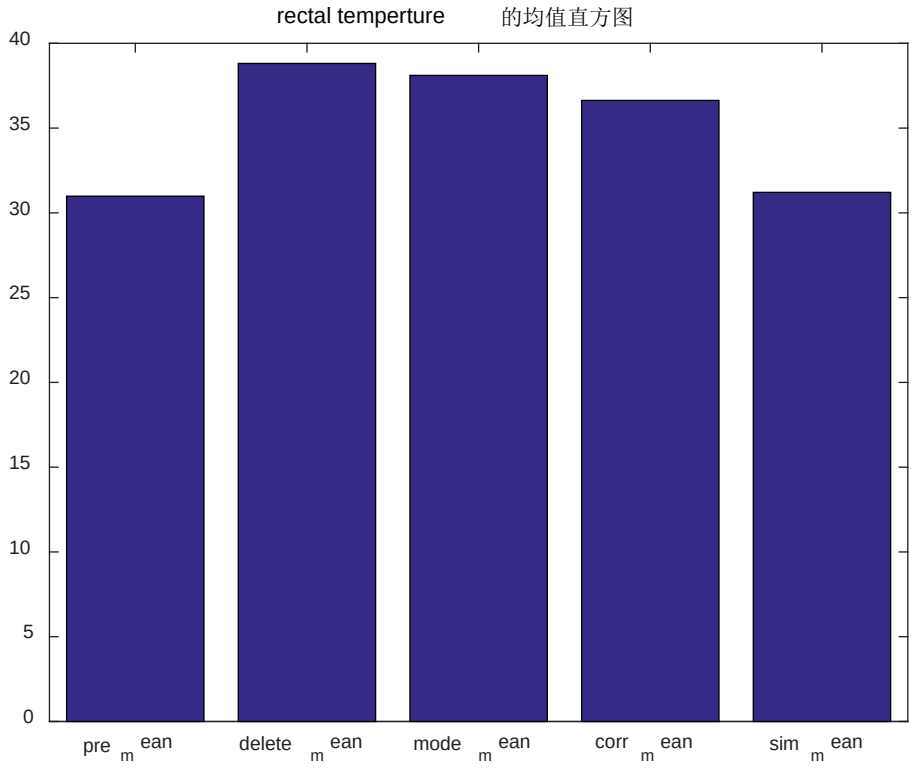
众数补齐后缺失数据：0

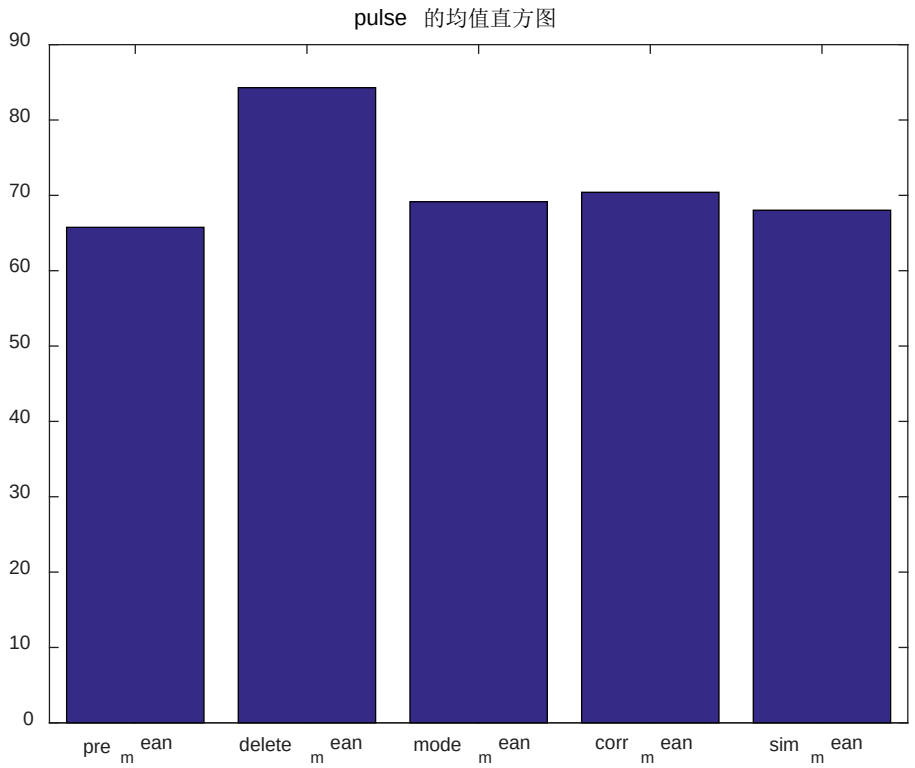
皮尔逊相关性补齐后缺失数据：1181

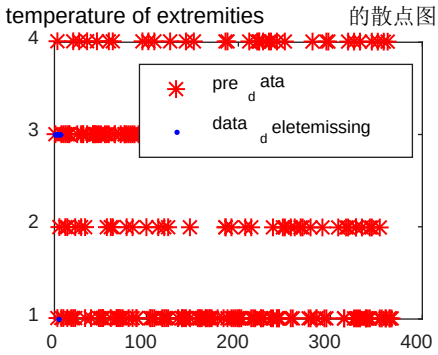
余弦相似度补齐后缺失数据：369

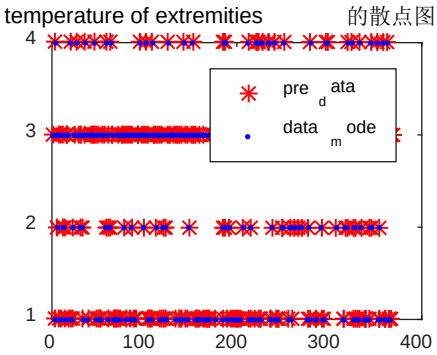
（4） 数据可视化比较（step8 输出结果）

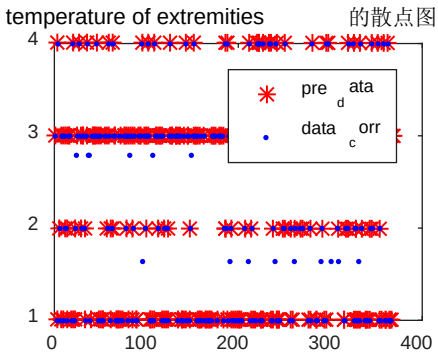
下面给出部分属性的比较结果

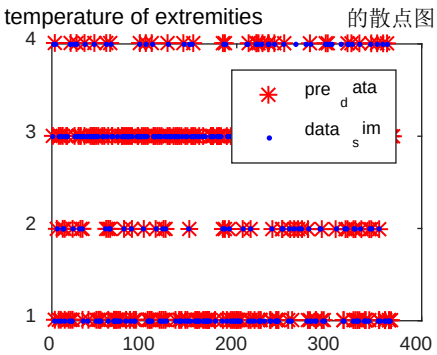

rectal temperture 的均值直方图
40
35
30
25
20
15
10
5
0
pre mean
delete mean
mode mean
corr mean
sim mean

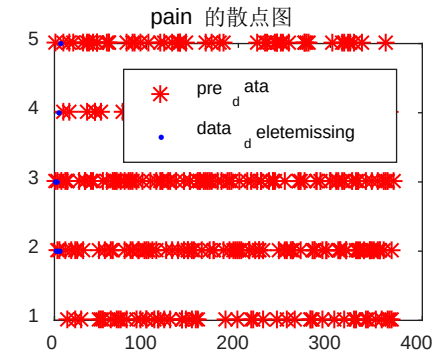

pulse 的均值直方图
90
80
70
60
50
40
30
20
10
0
pre mean
delete mean
mode mean
corr mean
sim mean

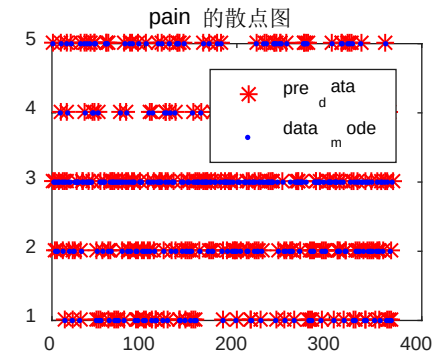


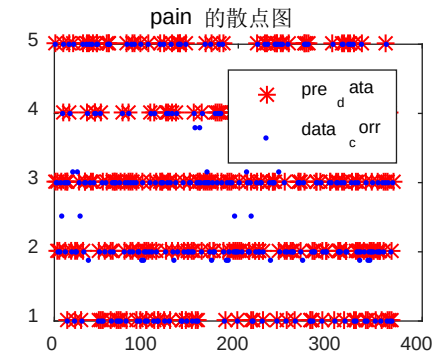


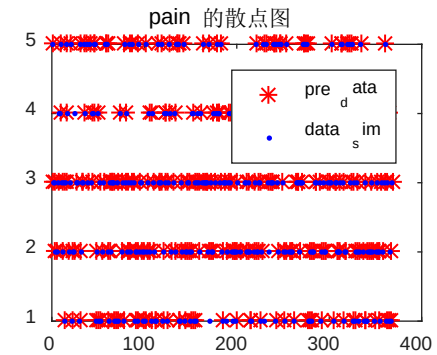












讨论：在对缺失值进行补充时，皮尔逊相关和余弦相似度测量得到的结果都没有很好的补齐缺失值，这是由于相关或相似的两个属性同时缺失，无法得到补充。