数据挖掘作业一

**—**马的疝病分析

2120161081

赵冬迪

2017/4/16

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。

**二、数据说明**

共368个样本，27个特征。

**三、数据分析要求**

1. 数据可视化和摘要

(1) 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。

(2) 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

对7种海藻，分别绘制其数量与标称变量，如size的条件盒图

2. 数据缺失的处理

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。

**四、实验环境**

Matlab R2016a

**五、实现方法**

**5.1 数据摘要**

**5.1.1 概述**

数据摘要部分主要进行的处理有以下两个方面：

1. 对标称属性，给出每个可能取值的频数；

2. 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。

**5.1.2 实现方法**

**（1）. 把原始数据导入excel表格**

通过excel表格的“数据”-->“来自文本” 导入

**（2）. 利用Matlab脚本实现数据摘要的获得**

脚本为：Data\_Abstract.m ，该脚本涉及到的关键函数如下：

tabulate(); %统计标称属性的词频并显示

max(); %统计数值属性的最大值

min(); %统计数值属性的最小值

mean(); %统计数值属性的均值

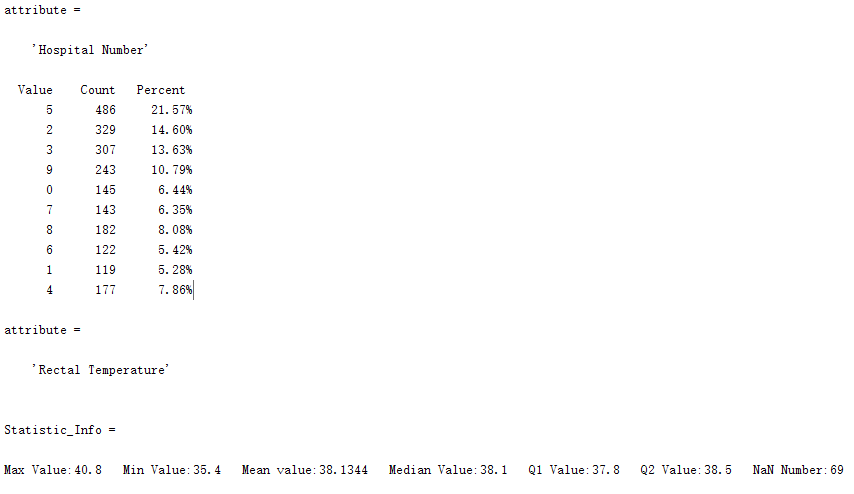
median(); %统计数值属性的中位数

prctile(col,25); %统计数值属性的上四分位数

prctile(col,75); 统计数值属性的下四分位数

**5.1.3 实验结果**

实验结果如下（分别选择一个数值型属性和标称型属性为例）：



**5.2 数据可视化**

**5.2.1 概述**

数据可视部分主要进行的处理是把数值部分的每一个属性包含的数据分别可视化（绘制成直方图，qq图，盒图）：

**5.2.2 核心代码**

%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

list = [4,5,6,16,19,20,22];

for i = 1:28

if ismember(i, list)

attribute = index(i);

col = data(:, i);

figure(find(list == i));

subplot(1,3,1), hist(col), title(index(i)); %绘制直方图

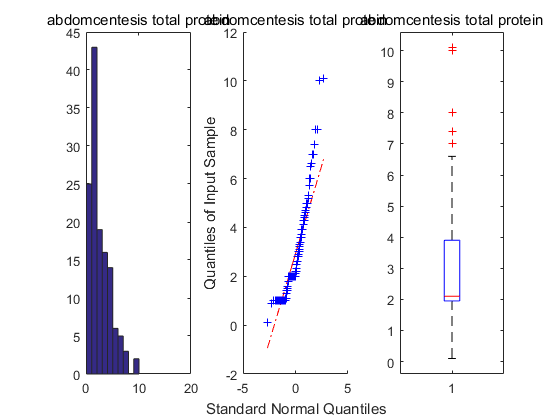
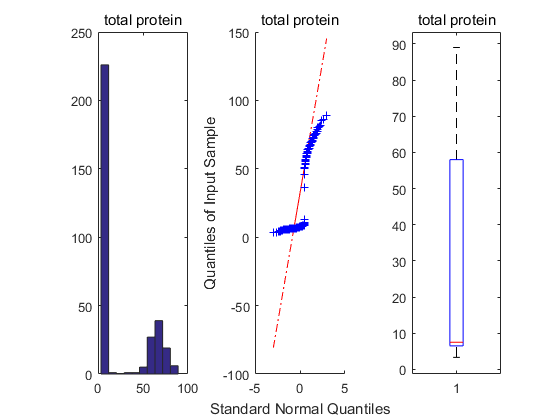
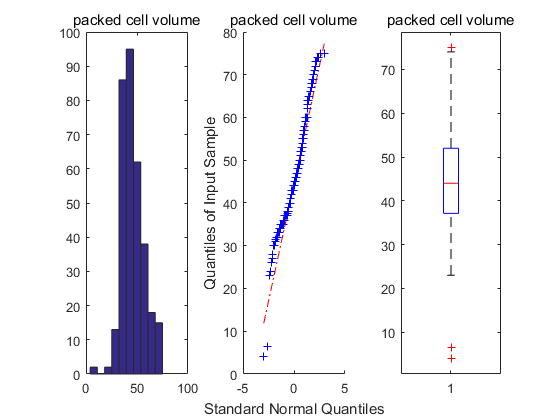
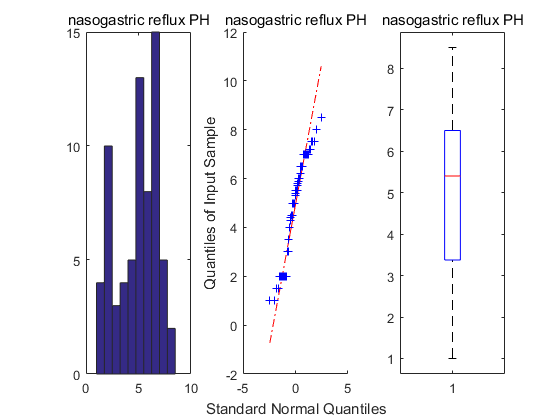
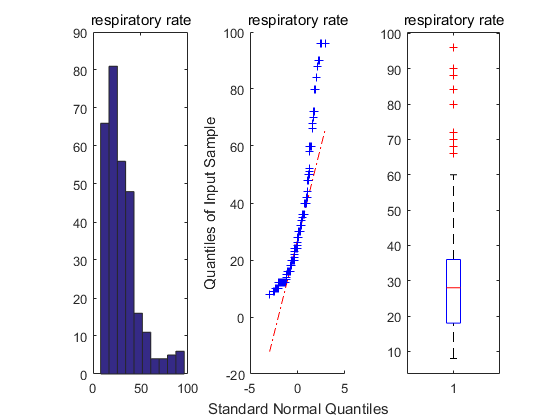
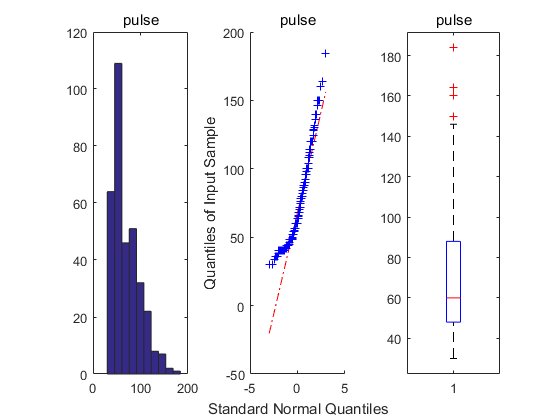
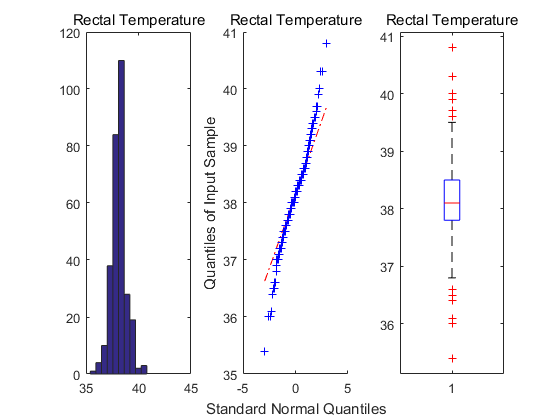
subplot(1,3,2), qqplot(col), title(index(i)); %绘制QQ图

subplot(1,3,3), boxplot(col), title(index(i)); %绘制盒图

end

end

**5.2.3 实验结果**

****

**5.2.3 结果分析**

**（1）绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。**

根据qq图，可以看出，属性‘Rectal Temperature’和‘packed cell volume’比较接近正态分布，其他属性均不太符合正态分布。通过直方图也可以验证这个结果。

**（2）绘制盒图，对离群值进行识别。**

由于现实数据中总是存在各式各样地“脏数据”，也成为“离群点”，于是为了不因这些少数的离群数据导致整体特征的偏移，将这些离群点单独汇出，而盒图中的两级修改成最小观测值与最大观测值。这里有个经验，就是最大(最小)观测值设置为与四分位数值间距离为1.5个IQR(中间四分位数极差)。即

IQR = Q3-Q1，即上四分位数与下四分位数之间的差，也就是盒子的长度。

最小观测值为min = Q1 - 1.5\*IQR，如果存在离群点小于最小观测值，则下限为最小观测值，离群点单独以点汇出。如果没有比最小观测值小的数，则下限为最小值。

最大观测值为max = Q3 +1.5\*IQR，如果存在离群点大于最大观测值，则上限为最大观测值，离群点单独以点汇出。如果没有比最大观测值大的数，则上限为最大值。

通过盒图，在分析数据的时候，盒图能够有效地帮助我们识别数据的特征：

（1）直观地识别数据集中的异常值(查看离群点)。

（2）判断数据集的数据离散程度和偏向。

在本实验中，可以通过这样的方法对数据离群点进行识别。

**5.3. 数据缺失的处理**

**1） 将缺失部分剔除**

剔除缺失数据与写入文件的核心代码如下：

%% 将缺失部分剔除

if option == 1

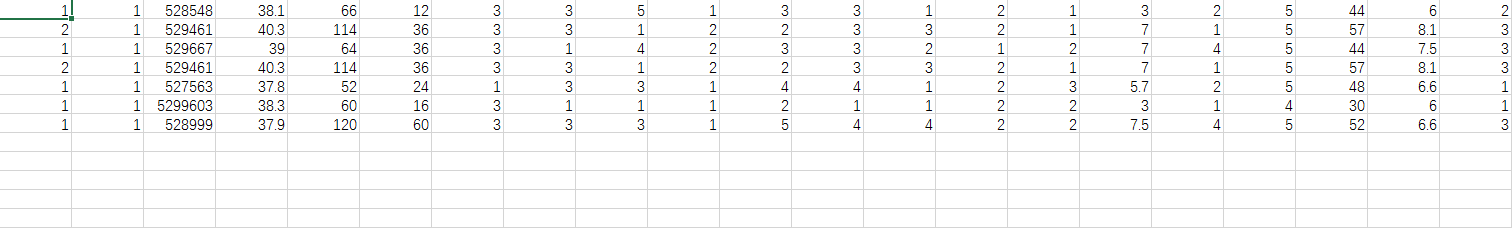
data(any(isnan(data),2),:) = [];

xlswrite('MissingValueProcessFile1.xlsx', data);

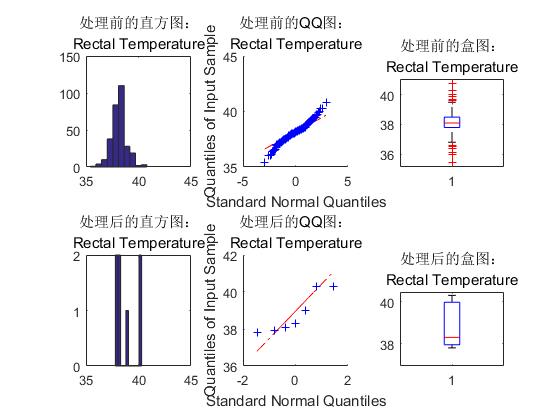
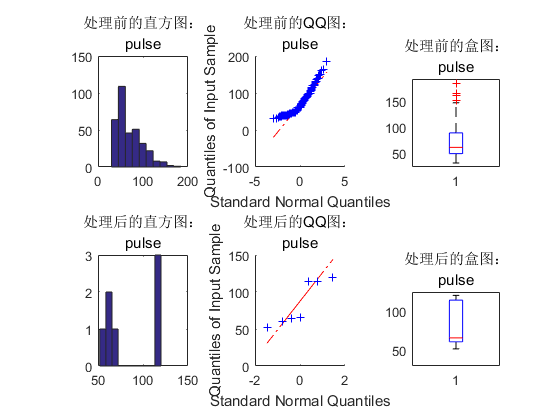
Data\_Compare(origin\_data, data);

end

处理前后数据如下：



可视化对比如下(举例说明)：



可以看出，通过这个方法处理后的数据只剩下7条，变得非常稀疏。因此对于这个数据集来说，这不是一个很好的处理方法。

**2) 使用高频数值来填补缺失值**

核心代码如下：

%% 用最高频率值来填补缺失值

if option == 2

for i = 1:28

hf = mode(data(:, i));

col = data(:, i);

pos = find(isnan(col));

data(pos, i) = hf;

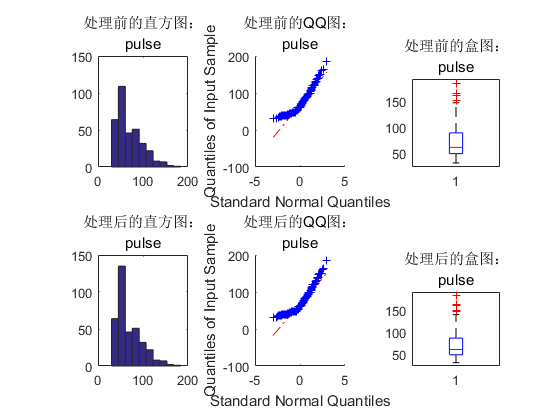
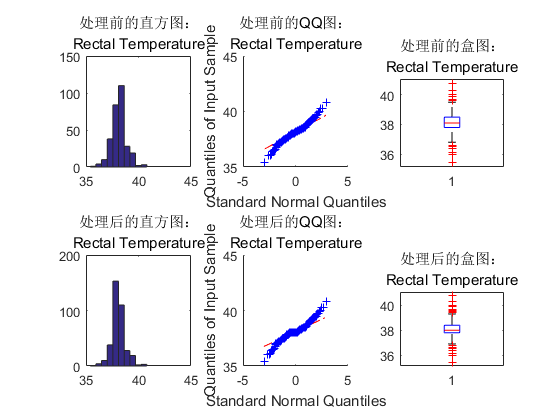
xlswrite('MissingValueProcessFile2.xlsx', data);

Data\_Compare(origin\_data, data);

end

end

处理后数据可视化对比如下：

****

**3) 通过属性的相关关系来填补缺失值**

核心代码如下：

%% 通过属性的相关关系来填补缺失值

if option == 3

cor\_mat = correlation\_mat\_attribute(data);

cor\_size = size(cor\_mat, 1); % 矩阵大小，正常情况下是方阵

for i = 1: m

for j = ATTRIBUTE\_L: ATTRIBUTE\_H

if(isnan(data(i, j)))

[~, index] = sort(cor\_mat(j - ATTRIBUTE\_L + 1, :));

index\_list = fliplr(index); % sort升序，fliplr翻转，变成降序，得到参考的属性优先度列表

flag = 0; % 标识是否补全成功

for k = 1: cor\_size

ref\_attr = index\_list(k); % 用于补全参考的属性

if(~isnan(data(i, ref\_attr)))

data(i, j) = standard\_line(j - ATTRIBUTE\_L + 1) / standard\_line(ref\_attr) \* ...

data(i, ref\_attr + ATTRIBUTE\_L - 1); % 按比例补全（这不是最好的方法）

flag = 1;

break;

end

end

if(flag == 0)

disp(['Insert fail at row ', num2str(i), ' col ', num2str(j)]);

return ;

end

end

end

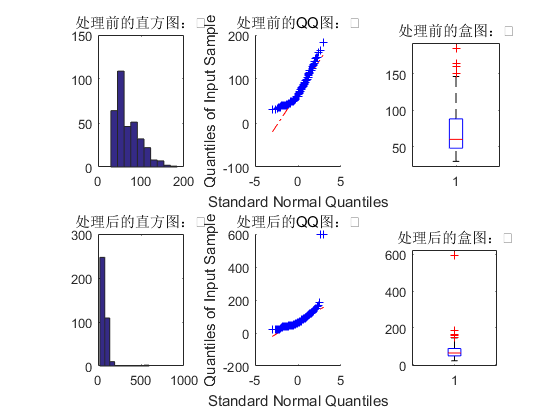
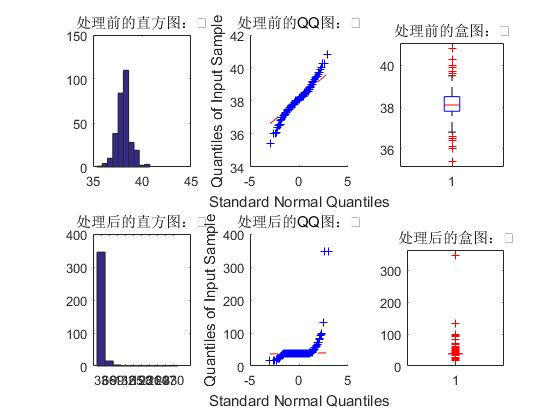
end

xlswrite('MissingValueProcessFile3.xlsx', data);

Data\_Compare(origin\_data, data);

end

处理后数据可视化对比如下（举例说明）：

****

**4） 使用数据对象之间的相似型填补缺失值**

%% 通过数据对象之间的相似性来填补缺失值

if option == 4

sim\_mat = similarity\_mat\_sample(data); % 获得相关性矩阵，函数见下

sim\_size = size(sim\_mat, 1); % 矩阵大小，正常情况下是方阵

for i = 1: m

for j = ATTRIBUTE\_L: ATTRIBUTE\_H

if(isnan(data(i, j)))

[~, index\_list] = sort(sim\_mat(i, :));

flag = 0; % 标识是否补全成功

for k = 1: sim\_size

ref\_samp = index\_list(k); % 用于补全参考的属性

if(~isnan(data(ref\_samp, j)))

data(i, j) = data(ref\_samp, j); % 原样填上，补全

flag = 1;

break;

end

end

if(flag == 0)

disp(['Insert fail at row ', num2str(i), ' col ', num2str(j)]);

return ;

end

end

end

end

xlswrite('MissingValueProcessFile4.xlsx', data);

Data\_Compare(origin\_data, data);

end

处理后数据可视化对比如下（举例说明）：

