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 2](#_Toc480220810)

[2. 数据分析要求 2](#_Toc480220811)

[3. 算法实现 2](#_Toc480220812)

[step1 读入数据 2](#_Toc480220813)

[step2 数据摘要 3](#_Toc480220814)

[step3 数据可视化 4](#_Toc480220815)

[step4 剔除缺失部分 4](#_Toc480220816)

[step5 用最高频率值填补缺失值 4](#_Toc480220817)

[step6 通过属性的相关关系来填补缺失值 5](#_Toc480220818)

[step7 通过数据对象之间的相似性来填补缺失值 6](#_Toc480220819)

[step8 可视化比较新旧数据集 7](#_Toc480220820)

[4. 结果与讨论 8](#_Toc480220821)

[（1） 数据摘要（step2输出结果） 8](#_Toc480220822)

[（2） 数据可视化（step3输出结果） 14](#_Toc480220823)

[（3） 数据缺失值处理（step4~step7输出结果） 18](#_Toc480220824)

[（4） 数据可视化比较（step8输出结果） 18](#_Toc480220825)

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。

1. 数据分析要求
2. 数据可视化和摘要
3. 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。

1. 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

1. 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。

1. 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

1. 结果与讨论
2. 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

1. 数据可视化（step3输出结果）











*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。

1. 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

1. 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









***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。