# 马的疝病分析过程报告

## 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。

## 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

以下是对17个标称属性频数的统计。频次一列的n个正整数分别对应左侧标称属性括号内1-n种属性的频次。

|  |  |
| --- | --- |
| 标称属性 | 频次 |
| surgery（1-2） | 214 152 |
| age（1-2） | 340 28 |
| temperature of extremities（1-4） | 95 39 135 34 |
| peripheral pulse（1-4） | 151 6 116 12 |
| mucous membranes（1-6） | 98 38 81 50 28 25 |
| capillary refill time（1-3） | 232 96 2 |
| pain（1-5） | 49 77 82 47 50 |
| peristalsis（1-4） | 49 22 154 91 |
| abdominal distension（1-4） | 101 75 85 42 |
| nasogastric tube（1-3） | 89 121 27 |
| nasogastric reflux（1-3） | 141 45 49 |
| rectal examination-feces（1-4） | 68 14 61 97 |
| abdomen（1-5） | 31 24 19 55 96 |
| abdominocentesis appearance（1-3） | 52 62 60 |
| outcome（1-3） | 225 89 52 |
| surgical lesion（1-2） | 232 136 |
| cp data（1-2） | 124 244 |

其实25-27个属性也是标称属性，但是因为数据的格式会产生歧义（例如，11110可以理解为11,1,1,0，也可以理解为1,1,1,10），所以在这里不做统计。

#### 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

以下为数值属性的最大、最小、均值、中位数、四分位数及缺失值的个数。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | max | min | mean | median | Q1 | Q3 | NaN |
| rectal temperature | 40.8 | 35.4 | 38.1 | 38.1 | 37.8 | 38.5 | 69 |
| pulse | 184 | 30 | 70.8 | 60 | 48 | 88 | 26 |
| respiratory rate | 96 | 8 | 30.5 | 28 | 18 | 36 | 71 |
| nasogastric reflux PH | 8.5 | 1 | 4.96 | 5.4 | 3.38 | 6.5 | 299 |
| packed cell volume | 75 | 4 | 45.7 | 44 | 37.1 | 52 | 37 |
| total protein | 89 | 3.3 | 24.8 | 7.5 | 6.5 | 58 | 43 |
| abdomcentesis total protein | 10.1 | 0.1 | 2.95 | 2.1 | 1.95 | 3.9 | 235 |

### 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

****

上述14张图分别是7种数值属性的直方图和qq图，除了total protein这个属性以外，其他的数据都比较服从正态分布。

#### 绘制盒图，对离群值进行识别

   

如图所示，7个数值属性的盒图，红色“+”表示离群点。

## 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

首先我们先去除掉所有样本28个属性中≥10个为NaN值的数据进行清除，清除了34条（12.23%）数据。



如图所示，以数值属性4（rectal temperature）为例，data为原数据，data1为data经过3.1处理的数据，效果在直方图上不是很明显，不过一些粗的空白bar的确消失了，而且数据量也只减小了12.23%。

### 用最高频率值来填补缺失值

其次，我们对data1数据中，属性损失率≤10%的样本，用最大频数的数据（标称属性）或者中位数（数值数据）来代替。





如图所示，以标称属性7（temperature of extremities），数值属性19（packed cell volume）为例，在缺失值很少的情况下，最大频率和中位数可以在不影响数据分布的情况下很好地补充空白数据。

### 通过属性的相关关系来填补缺失值

通过我们的观察，发现相关的属性只有14（nasogastric tube）和15（nasogastric reflux），他们都是标称属性，而且取值范围都是1-3。我们的做法是，按照组合出现的频数补充NaN（例如一个样本是[3 NaN]，我们就寻找[3 1]、[3 2]和[3 3]中出现最多次数的组合，并把[3 NaN]赋值为那个组合，对于[NaN,NaN]，我们就用9中组合中频次最大的那个给它赋值）



如图所示，以标称属性14（nasogastric tube）为例，空值不是简单地被最大频数填充，而是根据标称属性15（nasogastric reflux）的值予以补充。

### 通过数据对象之间的相似性来填补缺失值

对象之间的相似性，我们用二范数的倒数来表示。我们首先找到data3中还依旧存在NaN的属性，将所有样本中这些属性都不为空的样本作为标准，计算要填充样本与他们的距离，寻找出最短的，将NaN补充。

由于标准集中的元素过少，所以我们每填充一个样本，就把它放到标准集中，防止样本过于单一。



如图所示，以数值特征4（rectal temperature）为例，可以看到对data3中的数据进行了补充，使得整体变化不大。

### 处理后，可视化地对比新旧数据集

数值特征4（rectal temperature）的直方图和QQ图



标量特征18（abdomen）的直方图



通过观察上述图，结合四种缺失数据处理方法，缺失数据得到了很好地补充。