洪辉婷 2120160997

数据挖掘课程作业一实验报告

——马的疝病分析

一、 实验数据

医院对马的疝病(马胃肠痛)分析所需要的检测指标所组成的数据集。共 368 个样本，28 个特征（其中有一个特征为编号）。

二、 实验内容

根据所给的数据集：

（1） 分析该数据集的各个属性，其中;

- 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；绘制 qq 图；绘制盒图。

（2） 分别使用四种策略处理数据缺失值：

- 将缺失部分剔除：

  剔除掉含有缺失属性值的样本。剔除后，数据集只剩下 7 个数据完整的样本。

- 用最高频率值来填补缺失值：

  计算数据集中 27 个属性（包括标称属性和数值属性）的最高频率值，并用于填补对应属性的缺失值。

- 通过属性的相关关系来填补缺失值：

考虑 3 个比较重要的标称属性（2：age 年龄，7：temperature of extremities 四肢温度，13：abdominal distension 腹部膨胀）下的属性相关关系。由于属性较多且类型复杂，不方便画图分析各个属性的相关性分析，转而利用条件概率来进行相关性判断，条件概率公式如下：

$$\mathrm{P}(\theta_i|\theta_2,\theta_7,\theta_{13}) \coloneqq \frac{N(\theta_i,\theta_2,\theta_7,\theta_{13})}{N(\theta_2,\theta_7,\theta_{13})}$$

但由于该数据集太小，且属性缺失情况复杂，很难根据上面的公式得到统计结果，因此在本实验中，对上面的公式放宽条件，改为如下公式：

$$\mathrm{P}(\theta_i|\theta_2,\theta_7,\theta_{13}) \approx \frac{N(\theta_i,(\theta_2||\theta_7||\theta_{13}))}{N(\theta_2||\theta_7||\theta_{13})}$$

其中$N(\theta_2||\theta_7||\theta_{13})$表示第二个属性值为$\theta_2$或第七个属性值为$\theta_7$或第十三个个属性值为$\theta_{13}$的样本集合$S_{\theta_2||\theta_7||\theta_{13}}$的样本数量；$N(\theta_i,(\theta_2||\theta_7||\theta_{13}))$表示集合$S_{\theta_2||\theta_7||\theta_{13}}$中第 i 个属性值为$\theta_i$的样本数量。

最后，选取$\mathrm{P}(\theta_i|\theta_2,\theta_7,\theta_{13})$最高的属性值来填补该样本第 i 个属性缺失值。

- 通过数据对象之间的相似性来填补缺失值：

  利用 K 最近邻法（K=1）来填补缺失值。由于数据集中存在所有的数值属性都缺失的样本，所以使用所有的数值型属性（4，5，6，16，20，22）和 6 个所有样本都有值的标称属性（2，24，25，26，27，28）共同来进行样本

之间的距离度量。其中，标称属性使用 Hamming 距离，而数值型属性用欧式距离（的平方）。

但由于并不是所有的样本都有完整的数值型属性的值，因此我们将数值型属性的距离做平均处理。

三、 附件：实验程序及结果

1. 文件夹：原始数据

horse-colic.txt：原始数据集

horse_colic.py：分析数据集中各个属性的 python 程序

属性统计.png：各个属性的统计程序运行结果

属性统计 2.png：各个属性的统计程序运行结果（续）

histAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的直方图

histAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的直方图

histPacked cell volume.jpg：属性 19：packed cell volume 的直方图

histPulse.jpg：属性 5：pulse 的直方图

histRectal temperature.jpg：属性 4：rectal temperature 的直方图

histRespiratory rate.jpg：属性 6：respiratory rate 的直方图

histTotal protein.jpg：属性 20：total protein 的直方图

qqAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的 qq 图

qqAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的 qq 图

qqPacked cell volume.jpg：属性 19：packed cell volume 的 qq 图

qqPulse.jpg：属性 5：pulse 的 qq 图

qqRectal temperature.jpg：属性 4：rectal temperature 的 qq 图

qqRespiratory rate.jpg：属性 6：respiratory rate 的 qq 图

qqTotal protein.jpg：属性 20：total protein 的 qq 图

boxAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的盒图

boxAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的盒图

boxPacked cell volume.jpg：属性 19：packed cell volume 的盒图

boxPulse.jpg：属性 5：pulse 的盒图

boxRectal temperature.jpg：属性 4：rectal temperature 的盒图

boxRespiratory rate.jpg：属性 6：respiratory rate 的盒图

boxTotal protein.jpg：属性 20：total protein 的盒图

2. 文件夹：剔除缺失数据

horse-colic.txt：原始数据集

horse_colic_dropMissingData.py：剔除原始数据集中缺失数据的 python 程序

horse-colic-dropMissingData.txt：剔除缺失数据后得到的数据集

horse_colic.py：分析数据集 horse-colic-dropMissingData.txt 中各个属性的 python 程序

属性统计.png：各个属性的统计程序运行结果

histAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的直方图

histAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的直方图

histPacked cell volume.jpg：属性 19：packed cell volume 的直方图

histPulse.jpg：属性 5：pulse 的直方图

histRectal temperature.jpg：属性 4：rectal temperature 的直方图

histRespiratory rate.jpg：属性 6：respiratory rate 的直方图

histTotal protein.jpg：属性 20：total protein 的直方图

qqAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的 qq 图

qqAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的 qq 图

qqPacked cell volume.jpg：属性 19：packed cell volume 的 qq 图

qqPulse.jpg：属性 5：pulse 的 qq 图

qqRectal temperature.jpg：属性 4：rectal temperature 的 qq 图

qqRespiratory rate.jpg：属性 6：respiratory rate 的 qq 图

qqTotal protein.jpg：属性 20：total protein 的 qq 图

boxAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的盒图

boxAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的盒图

boxPacked cell volume.jpg：属性 19：packed cell volume 的盒图

boxPulse.jpg：属性 5：pulse 的盒图

boxRectal temperature.jpg：属性 4：rectal temperature 的盒图

boxRespiratory rate.jpg：属性 6：respiratory rate 的盒图

boxTotal protein.jpg：属性 20：total protein 的盒图

3. 文件夹：最高频率填补缺失值

horse-colic.txt：原始数据集

horse_colic_fillWithMostFrequency.py：用最高频率来填补原始数据中的缺失值的 python 程序

horse-colic-fillWithMostFrequency.txt：填补缺失数据后得到的数据集

horse_colic.py：分析数据集 horse-colic-fillWithMostFrequency.txt 中各个属性的 python 程序

属性统计.png：各个属性的统计程序运行结果

属性统计 2.png：各个属性的统计程序运行结果（续）

histAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的直方图

histAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的直方图

histPacked cell volume.jpg：属性 19：packed cell volume 的直方图

histPulse.jpg：属性 5：pulse 的直方图

histRectal temperature.jpg：属性 4：rectal temperature 的直方图

histRespiratory rate.jpg：属性 6：respiratory rate 的直方图

histTotal protein.jpg：属性 20：total protein 的直方图

qqAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的 qq 图

qqAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的 qq 图

qqPacked cell volume.jpg：属性 19：packed cell volume 的 qq 图

qqPulse.jpg：属性 5：pulse 的 qq 图

qqRectal temperature.jpg：属性 4：rectal temperature 的 qq 图

qqRespiratory rate.jpg：属性 6：respiratory rate 的 qq 图

qqTotal protein.jpg：属性 20：total protein 的 qq 图

boxAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的盒图

boxAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的盒图

boxPacked cell volume.jpg：属性 19：packed cell volume 的盒图

boxPulse.jpg：属性 5：pulse 的盒图

boxRectal temperature.jpg：属性 4：rectal temperature 的盒图

boxRespiratory rate.jpg：属性 6：respiratory rate 的盒图

boxTotal protein.jpg：属性 20：total protein 的盒图

4. 文件夹：属性相关性填补缺失值

horse-colic.txt：原始数据集

horse_colic_fillWithRelavity.py：利用属性相关性来填补原始数据中的缺失值的 python 程序

horse-colic-fillWithRelativity.txt：填补缺失数据后得到的数据集

horse_colic.py：分析数据集 horse-colic-fillWithRelativity.txt 中各个属性的 python 程序

属性统计.png：各个属性的统计程序运行结果

属性统计 2.png：各个属性的统计程序运行结果（续）

histAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的直方图

histAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的直方图

histPacked cell volume.jpg：属性 19：packed cell volume 的直方图

histPulse.jpg：属性 5：pulse 的直方图

histRectal temperature.jpg：属性 4：rectal temperature 的直方图

histRespiratory rate.jpg：属性 6：respiratory rate 的直方图

histTotal protein.jpg：属性 20：total protein 的直方图

qqAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的 qq 图

qqAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的 qq 图

qqPacked cell volume.jpg：属性 19：packed cell volume 的 qq 图

qqPulse.jpg：属性 5：pulse 的 qq 图

qqRectal temperature.jpg：属性 4：rectal temperature 的 qq 图

qqRespiratory rate.jpg：属性 6：respiratory rate 的 qq 图

qqTotal protein.jpg：属性 20：total protein 的 qq 图

boxAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的盒图

boxAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的盒图

boxPacked cell volume.jpg：属性 19：packed cell volume 的盒图

boxPulse.jpg：属性 5：pulse 的盒图

boxRectal temperature.jpg：属性 4：rectal temperature 的盒图

boxRespiratory rate.jpg：属性 6：respiratory rate 的盒图

boxTotal protein.jpg：属性 20：total protein 的盒图

5. 文件夹：属性相似性填补缺失值

horse-colic.txt：原始数据集

count.py：统计各个属性缺失数量的 python 程序

属性缺失统计.png：count.py 程序的输出结果

horse_colic_fillWithSimilarity.py：利用属性相似性来填补原始数据中的缺失值的 python 程序

horse-colic-fillWithSimilarity.txt：填补缺失数据后得到的数据集

horse_colic.py：分析数据集 horse-colic-fillWithSimilarity.txt 中各个属性的 python 程序

属性统计.png：各个属性的统计程序运行结果

属性统计 2.png：各个属性的统计程序运行结果（续）

histAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的直方图

histAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的直方图

histPacked cell volume.jpg：属性 19：packed cell volume 的直方图

histPulse.jpg：属性 5：pulse 的直方图

histRectal temperature.jpg：属性 4：rectal temperature 的直方图

histRespiratory rate.jpg：属性 6：respiratory rate 的直方图

histTotal protein.jpg：属性 20：total protein 的直方图

qqAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的 qq 图

qqAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的 qq 图

qqPacked cell volume.jpg：属性 19：packed cell volume 的 qq 图

qqPulse.jpg：属性 5：pulse 的 qq 图

qqRectal temperature.jpg：属性 4：rectal temperature 的 qq 图

qqRespiratory rate.jpg：属性 6：respiratory rate 的 qq 图

qqTotal protein.jpg：属性 20：total protein 的 qq 图

boxAbdomcentesis total protein.jpg：属性 22：Abdomcentesis total protein 的盒图

boxAsogastric reflux PH.jpg：属性 16：nasogastric reflux PH 的盒图

boxPacked cell volume.jpg：属性 19：packed cell volume 的盒图

boxPulse.jpg：属性 5：pulse 的盒图

boxRectal temperature.jpg：属性 4：rectal temperature 的盒图

boxRespiratory rate.jpg：属性 6：respiratory rate 的盒图

boxTotal protein.jpg：属性 20：total protein 的盒图