Q3 实验报告

1854116 朱明志

目录

1 数据描述

本实验的数据集来源于 OCR 识别视频帧产生的日志，视频是某定区公交车司机处摄像头拍下的监控视频录像。该视频为每秒 12 帧的视频，本次截取了 5s 的视频生成了 OCR 识别日志。因为 OCR 识别并不能达到 100%成功，有时面对视频中复杂的背景噪声会出现无法识别的问题，因此日志中的某些数据的某些属性值存在缺失。具体日志数据已附在 data 子文件夹里。

视频截图： 

OCR 的操作是将图片中的经纬度坐标、速度值、时间日期信息识别出来并输出至日志 csv 文件保存，本次实验便使用了这个日志文件作为数据集。

2 代码运行结果屏幕拷贝

2.1 缺失值处理代码运行结果

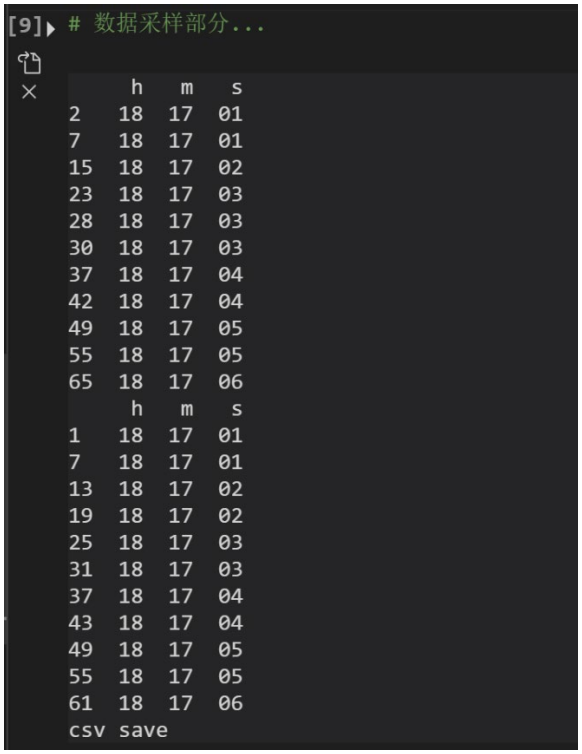
```
[10]# 导入一个pandas数据框架下...

     h   m   s
0   18  17  01
1   18  17  01
2   18  17  01
3   18  17  01
4   18  17  01
..  ..  ..  ..
61  18  17  06
62  18  17  06
63  18  17  06
64  18  17  06
65  18  17  06

[66 rows x 3 columns]
(66, 3)
[11]# 数据缺失值处理部分...

csv save
```

2.2 数据采样代码运行结果



3 实验内容讨论

3.1 缺失值处理

3.1.1 实验原理

处理不完整数据集的方法主要有三大类： 删除元组、数据补齐、不处理 。

- 删除元组

也就是将存在遗漏信息属性值的对象（元组，记录）删除，从而得到一个完备的信息表 。这种方法简单易行， 在对象有多个属性缺失值、被删除的含缺失值的对象与初始数据集的数据量相比非常小的情况下非常有效，类标号缺失时通常使用该方法。

然而，这种方法却有很大的局限性。 它以减少历史数据来换取信息的完备，会丢弃大量隐藏在这些对象中的信息。 在初始数据集包含的对象很少的情况下，删除少量对象足以严重影响信息的客观性和结果的正确性；因此，当缺失数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。删除元组，或者直接删除该列特征，有时候会导致性能下降。

- 数据补齐

这类方法是用一定的值去填充空值，从而使信息表完备化。通常基于统计学原理，根据初始数据集中其余对象取值的分布情况来对一个缺失值进行填充 。数据挖掘中常用的有以下几种补齐方法：

✧ 人工填写（filling manually）

由于最了解数据的还是用户自己，因此这个方法产生数据偏离最小，可能是填充效果最好的一种。然而一般来说，该方法很费时，当数据规模很大、空值很多的时候，该方法是不可行的。

✧ 特殊值填充（Treating Missing Attribute values as Special values）

将空值作为一种特殊的属性值来处理，它不同于其他的任何属性值 。如所有的空值都用“unknown”填充。这样将形成另一个有趣的概念， 可能导致严重的数据偏离，一般不推荐使用 。

✧ 平均值填充（Mean/Mode Completer）

将初始数据集中的属性分为数值属性和非数值属性来分别进行处理 。 如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值；如果空值是非数值型的，就根据统计学中的众数原理 ，用该属性在其他所有对象的取值次数最多的值(即出现频率最高的值)来补齐该缺失的属性值。与其相似的另一种方法叫条件平均值填充法（Conditional Mean Completer）。在该方法中，用于求平均的值并不是从数据集的所有对象中取，而是从与该对象具有相同决策属性值的对象中取得。

这两种数据的补齐方法，其基本的出发点都是一样的，以最大概率可能的取值来补充缺失的属性值，只是在具体方法上有一点不同。与其他方法相比，它是用现存数据的多数信息来推测缺失值。

✧ 热卡填充（Hot deck imputation，或就近补齐）

对于一个包含空值的对象，热卡填充法在完整数据中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。该方法概念上很简单，且利用了数据间的关系来进行空值估计。这个方法的缺点在于难以定义相似标准，主观因素较多。

✧ K 最近距离邻法（K-means clustering）

先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的 K 个样本，将这 K 个值加权平均来估计该样本的缺失数据。

✧ 使用所有可能的值填充（Assigning All Possible values of the Attribute）

用空缺属性值的所有可能的属性取值来填充，能够得到较好的补齐效果。但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大，可能的测试方案很多。

✧ 组合完整化方法（Combinatorial Completer）

用空缺属性值的所有可能的属性取值来试，并从最终属性的约简结果中选择最好的一个作为填补的属性值。这是以约简为目的的数据补齐方法，能够得到好的约简结果；但是，当数据量很大或者遗漏的属性值较多时，其计算的代价很大。

✧ 回归（Regression）

基于完整的数据集，建立回归方程。对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。当变量不是线性相关时会导致有偏差的估计。

✧ 期望值最大化方法（Expectation maximization，EM）

EM 算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。在每一迭代循环过程中交替执行两个步骤：E 步（Excepctaion step,期望步），在给定完全数据和前一次迭代所得到的参数估计的情况下计算完全数据对应的对数似然函数的条件期望；M 步（Maximzation step，极大化步），用极大化对数似然函数以确定参数的值，并用于下步的迭代。算法在 E 步和 M 步之间不断迭代直至收敛，即两次迭代之间的参数变化小于一个预先给定的阈值时结束。该方法可能会陷入局部极值，收敛速度也不是很快，并且计算很复杂。

就几种基于统计的方法而言，删除元组法和平均值法差于热卡填充法、期望值最大化方法和多重填充法；回归是比较好的一种方法，但仍比不上 hot deck 和 EM；EM 缺少 MI 包含的不确定成分。值得注意的是，这些方法直接处理的是模型参数的估计而不是空缺值预测本身。它们合适于处理无监督学习的问题，而对有监督学习来说，情况就不尽相同了。譬如，你可以删除包含空值的对象用完整的数据集来进行训练，但预测时你却不能忽略包含空值的对象。

3.1.2 实验过程

本次实验使用了数据集数据的 time 属性。

首先将 time 属性中的值提取出来，分割为小时、分钟、秒三列属性，可以明显看到，因为 OCR 识别成功率不能 100%的原因，有些行的某些数据是空值，需要对这些空值进行处理。（见 test.csv）

本次实验分别实验了四种不同的方法对这些缺失值进行了填补：

1. 用 nan 这个特殊值对空值进行填充（test1.csv）；
2. 直接将存在空值属性的数据删去（test2.csv）；
3. 当一个数据的某一个属性值为空时，使用在这个数据之前，最邻近这个数据的 1 条数据的这一属性值的平均值填充这个空值（test3.csv）；
4. 当一个数据的某一个属性值为空时，使用这个属性列中所有值的平均值填充这个空值（test4.csv）。

使用 Python 语言进行实现，具体实现见代码。

3.2 数据采样

3.2.1 实验原理

- 随机抽样：

  直接从整体数据中等概率抽取 n 个样本。这种方法优势是，简单、好操作、适用于分布均匀的场景；缺点是总体大时无法一一编号

- 系统抽样：

又称机械、等距抽样，将总体中个体按顺序进行编号，然后计算出间隔，再按照抽样间隔抽取个体。优势,易于理解、简便易行。缺点是，如有明显分布规律时容易产生偏差。

- 分层抽样：

先按照观察指标影响较大的某一种特征，将总体分若干个类别，再从每一层随机抽取一定数量的单位合并成总体。优点样本代表性好，少误差。

3.2.2 实验过程

本次实验使用了数据集数据的 time 属性。

首先利用上一个实验的成果，使用已经填补过缺失值的数据集作为抽样对象。考虑到 1 秒的视频足有 12 帧，OCR 识别和司机模式识别算法速度有限，为了提高实时响应速度，我只需要原有数据集的 1/6 大小即可，即 1 秒的图片帧中取样 2 帧作为代表跑 OCR 识别和模式识别算法。而 1 秒内司机的行为和显示时间不会有明显变化，这样的取样思想应该是合理且能有效提高识别效率的办法。

本次实验分别实验了三种不同的采样方法对这些数据进行了采样：

1. 随机采样（sample1.csv）；
2. 每 6 帧图片为一组，先分组再组内随机采样 1 帧图片（sample2.csv）；
3. 系统采样，采样距离为每间隔 6 帧图片采样一帧（sample3.csv）。

使用 Python 语言进行实现，具体实现见代码。

4 实验结果图表与分析

4.1 缺失值处理

表格文件名	表格内容
test.csv	未处理的 time 列提取数据表
test1.csv	使用特殊值填补的处理后数据表
test2.csv	使用删除元组方式处理后的数据表
test3.csv	使用 K-mean 填补法填补的数据表
test4.csv	使用平均值填补法填补的数据表

数据表格的原始文件都在 csv 子文件夹下，每个数据表的具体内容见上表。

下表展示了每一个含缺失值的数据经过 4 种不同方法处理后得到的数据：

原数据			特殊值填充			元组删除			K-mean填充			平均值填充		
h	m	s	h	m	s	h	m	s	h	m	s	h	m	s
18	17		18	17	nan				18	17	1	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18	17		18	17	nan				18	17	3	18	17	3
18			18	nan	nan				18	17	5	18	17	3
18	17		18	17	nan				18	17	5	18	17	3
18		5	18	nan	5				18	17	5	18	17	5

对于本数据而言，因为是日志文件，且日志内容缺失的是时间信息，所以用平均值填补法显然是不可行的，同理用特殊值填补也是不可行的。考虑到日志的特性，在一条日志生成时，下一条日志是什么完全不可预测，所以对 K-mean 填补法进行了修改，修改为该条数据以上最近的 k 条数据该属性值的平均值，又由于时间数据的连贯性，这个填补法有着最优的表现。最后，对于删除元组法，考虑到 1s 的视频有 12 帧，OCR 在绝大多数情况下不会对这 12 帧图片的时间全部识别失败，删去一些图片帧，日志的时间信息依然是连贯的，从这一点来说，这种处理方法也有一定是应用价值。

4.2 数据采样

表格文件名	表格内容
sample1.csv	随机采样
sample2.csv	分组随机采样
sample3.csv	系统采样

数据表格的原始文件都在 csv 子文件夹下，每个数据表的具体内容见上表。

下表展示了补全缺失值的数据经过 3 种不同采样方法处理后得到的子数据:

随机采样			分组随机采样			系统采样		
h	m	s	h	m	s	h	m	s
18	17	3	18	17	1	18	17	1
18	17	2	18	17	1	18	17	1
18	17	2	18	17	2	18	17	2
18	17	4	18	17	3	18	17	2
18	17	1	18	17	3	18	17	3
18	17	1	18	17	3	18	17	3
18	17	3	18	17	4	18	17	4
18	17	6	18	17	4	18	17	4
18	17	6	18	17	5	18	17	5
18	17	4	18	17	5	18	17	5
18	17	5	18	17	6	18	17	6

对于本数据而言，时间数据显然应该是连续的，而随机采样的结果破坏了这种时序性，这使得随机采样在本数据的处理中变的毫无意义。对于分组随机采样和系统采样而言，它们都很好的保留了这种时序性，都是可行的处理方法。但是分组随机采样还有一定的取样的随机性，使取样更能真实反应样本的数据特性，可以优先在数据处理中考虑使用。