数据探索性分析与预处理-海藻数据的分析

计算机科学与技术 李凯霞 2120151003 硕士

一、 问题描述

某些高浓度的有害藻类对河流生态环境的破坏是一个严重的问题。它们不仅破坏河流的生物，也破坏水质。能够监测并在早期对海藻的繁殖进行预测对提高河流质量是很有必要的。

针对这一问题的预测目标，在大约一年的时间内，在不同时间内收集了欧洲多条河流的水样。对于每个水样，测定了它们的不同化学性质以及 7 种有害藻类的存在频率。在水样收集过程中，也记录了一些其他特性，如收集的季节、河流大小和水流速度。

二、 数据说明

有 200 个水样，每条记录是同一条河流在该年的同一个季节的三个月内收集的水样的平均值。每条记录由 11 个变量构成，3 个是标称变量，分别描述水样收集的季节 season，河流大小 size 和河水速度 speed，剩下的 8 个变量是水样的化学参数，其中缺失值用 XXXXXXX 代替：

- 最大 pH 值(mxPH)
- 最小含氧量(mnO2)
- 平均氯化物含量(Cl)
- 平均硝酸盐含量(NO3)
- 平均氨含量(NH4)
- 平均正磷酸盐含量(oPO4)
- 平均磷酸盐含量(PO4)
- 平均叶绿素含量(Chla)

a1-a7 为 7 种不同有害藻类在相应水样中的频率数目。数据样本如下图:

autumn	small	medium	8.40000	9.90000	34.50000	2.81800	3515.00000	20.00000	47.00000	2.30000	13.60000	9.10000	0.00000	0.00000	1.40000	0.00000	0.00000
winter	small	medium	8.27000	7.80000	29.20000	0.05000	6400.00000	7.40000	23.00000	0.90000	5.30000	40.70000	3.30000	0.00000	0.00000	0.00000	1.90000
summer	small	medium	8.66000	8.40000	30.52300	3.44400	1911.00000	58.87500	84.46000	3.60000	18.30000	12.40000	1.00000	0.00000	0.00000	0.00000	1.00000
winter	small	high	8.30000	10.90000	1.17000	0.73500	13.50000	1.62500	3.00000	0.20000	66.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
spring	small	high	8.00000	XXXXXXX	1.45000	0.81000	10.00000	2.50000	3.00000	0.30000	75.80000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
winter	small	medium	8.30000	8.90000	20.62500	3.41400	228.75000	196.62000	253.25000	12.32000	2.00000	38.50000	4.10000	2.20000	0.00000	0.00000	10.20000
spring	small	medium	8.10000	10.50000	22.28600	4.07100	178.57001	182.42000	255.28000	8.95700	2.20000	2.70000	1.00000	3.70000	2.70000	0.00000	0.00000
winter	small	medium	8.00000	5.50000	77.00000	6.09600	122.85000	143.71001	296.00000	3.70000	0.00000	5.90000	10.60000	1.70000	0.00000	0.00000	7.10000
summer	small	medium	8.15000	7.10000	54.19000	3.82900	647.57001	59.42900	175.04601	13.20000	0.00000	0.00000	0.00000	5.70000	11.30000	17.00000	1.60000
winter	small	high	8.30000	7.70000	50.00000	8.54300	76.00000	264.89999	344.60001	22.50000	0.00000	40.90000	7.50000	0.00000	2.40000	1.50000	0.00000
spring	small	high	8.30000	8.80000	54.14300	7.83000	51.42900	276.85001	325.85699	11.84000	4.10000	3.10000	0.00000	0.00000	19.70000	17.00000	0.00000
winter	small	high	8.40000	13.40000	69.75000	4.55500	37.50000	10.00000	40.66700	3.90000	51.80000	4.10000	0.00000	0.00000	3.10000	5.50000	0.00000
spring	small	high	8.30000	12.50000	87.00000	4.87000	22.50000	27.00000	43.50000	3.30000	29.50000	1.00000	2.70000	3.20000	2.90000	9.60000	0.00000
autumn	small	high	8.00000	12.10000	66.30000	4.53500	39.00000	16.00000	39.00000	0.80000	54.40000	3.40000	1.20000	0.00000	18.70000	2.00000	0.00000
winter	small	low	XXXXXXX	12.60000	9.00000	0.23000	10.00000	5.00000	6.00000	1.10000	35.50000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
spring	small	medium	7.60000	9.60000	15.00000	3.02000	40.00000	27.00000	121.00000	2.80000	89.80000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
autumn	small	medium	7.29000	11.21000	17.75000	3.07000	35.00000	13.00000	20.81200	12.10000	24.80000	7.40000	0.00000	2.50000	10.60000	17.10000	3.20000
winter	small	medium	7.60000	10.20000	32.30000	4.50800	192.50000	12.75000	49.33300	7.90000	0.00000	0.00000	0.00000	4.60000	1.20000	0.00000	3.90000

图 1 数据样本示例

三、 数据分析

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 Max、最小 Min、均值 Mean、中位数 Median、四分位数 1st Qu/3rd Qu 及缺失值 NA 的个数。本实验数据摘要结果保存在..\result\summary.txt 中，如下图:

```
   season       size       speed        mxPH           mnO2             Cl              NO3
autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500   Min.   :  0.222   Min.   : 0.050
spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725   1st Qu.: 10.981   1st Qu.: 1.296
summer:45   small :71   medium:83   Median :8.060   Median : 9.800   Median : 32.730   Median : 2.675
winter:62                           Mean   :8.012   Mean   : 9.118   Mean   : 43.636   Mean   : 3.282
                                    3rd Qu.:8.400   3rd Qu.:10.800   3rd Qu.: 57.824   3rd Qu.: 4.446
                                    Max.   :9.700   Max.   :13.400   Max.   :391.500   Max.   :45.650
                                    NA's   :1       NA's   :2        NA's   :10        NA's   :2
     NH4                oPO4             PO4              Chla              a1              a2
Min.   :    5.00   Min.   :  1.00   Min.   :  1.00   Min.   :  0.200   Min.   : 0.00   Min.   : 0.000
1st Qu.:   38.33   1st Qu.: 15.70   1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50   1st Qu.: 0.000
Median :  103.17   Median : 40.15   Median :103.29   Median :  5.475   Median : 6.95   Median : 3.000
Mean   :  501.30   Mean   : 73.59   Mean   :137.88   Mean   : 13.971   Mean   :16.92   Mean   : 7.458
3rd Qu.:  226.95   3rd Qu.: 99.33   3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80   3rd Qu.:11.375
Max.   :24064.00   Max.   :564.60   Max.   :771.60   Max.   :110.456   Max.   :89.80   Max.   :72.600
NA's   :2          NA's   :2        NA's   :2        NA's   :12
      a3              a4              a5              a6              a7
Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
Median : 1.550   Median : 0.000   Median : 1.900   Median : 0.000   Median : 1.000
Mean   : 4.309   Mean   : 1.992   Mean   : 5.064   Mean   : 5.964   Mean   : 2.495
3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500   3rd Qu.: 6.925   3rd Qu.: 2.400
Max.   :42.800   Max.   :44.600   Max.   :44.400   Max.   :77.600   Max.   :31.600
```

图 2 数据摘要运行结果

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。Q-Q 图绘制变量值和正态分布的理论分位数（实线）的散点图，同时给出正态分布的 95%置信区间的带状图（虚线）。

在本实验中，通过直方图观察 mxPH 非常符合正态分布，通过 Q-Q 图检验，其数据点大多数都在 95%置信区间内。通过直方图观察 NH4 不符合正态分布，通过 Q-Q 图检验，其数据点大多数都在 95%置信区间外。同样观察得，mnO2 比较符合正态分布，其余参数均不符合。mxPH、NH4 的直方图及 Q-Q 图如下。

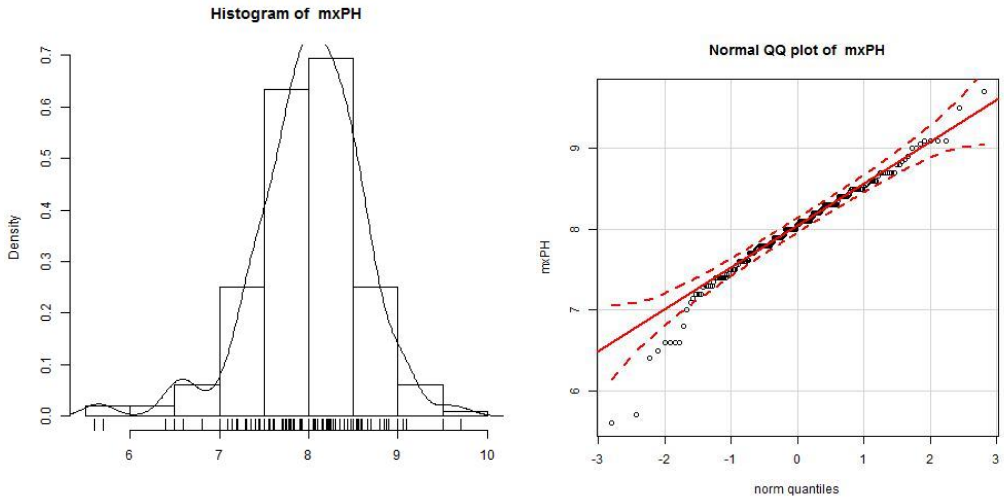

图 3 变量 mxPH 的直方图和 Q-Q 图

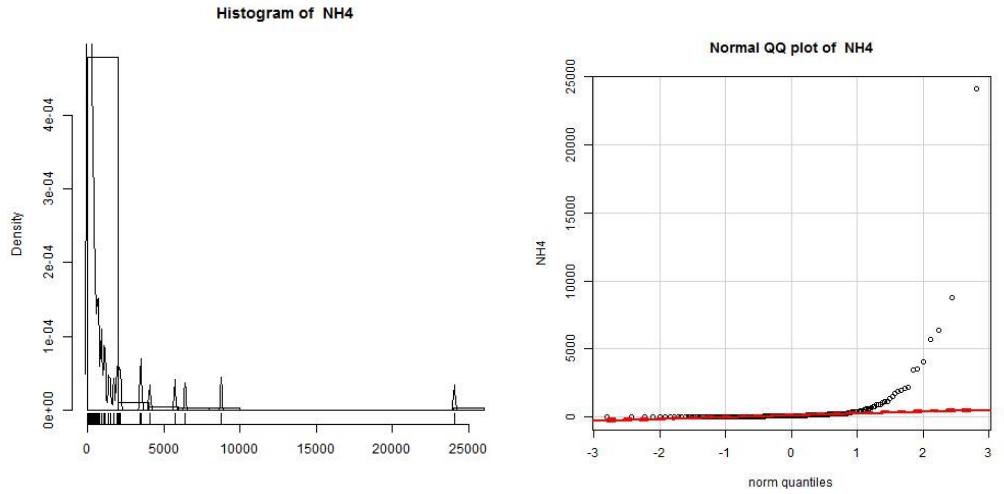

图 4 变量 NH4 的直方图和 Q-Q 图

绘制盒图，对离群值进行识别。盒图的边界代表变量的第一个四分位数和第三个四分位数，而框内的水平线是变量的中位数。设 r 是变量的四分位距，盒图上方的小横线是小于或等于第三个四分位数加 1.5*r 的最大的观测值，而盒图下方的小横线是大于或等于第一个四分位数减去 1.5*r 的最小观测值。盒图上方小横线上面或者下方小横线下面的数据点通常认为是离群点。mn02 的盒图如下。

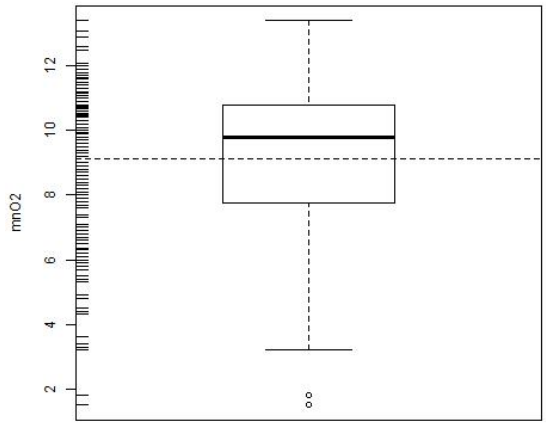

图 5 变量 mnO2 的盒图

对 7 种海藻，分别绘制其数量与标称变量，如 size 的条件盒图。条件绘图是依赖于某个特定因子的图形表示。因子是一个取值为有限集合的标称变量。例如，对于标称变量 size 的某个特定值的样本子集，可以研究标称变量 size 如何影响变量 a1 值的分布。海藻 a1 基于 size 的条件盒图如下。

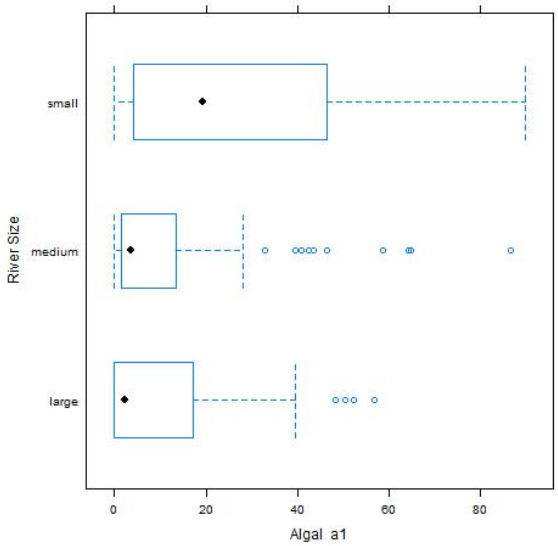

图 6 海藻 a1 基于 size 的条件盒图

为了便于对比，本实验原始数据输出为..\result\Analysis_Orginal.csv 文件中。可视化结果在..\result\picture\Orginal 中

3、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:

(1) 将缺失部分剔除

剔除缺失值，优点是实现简单，当缺失记录所占比例在可用数据集中非常小的时候，可以选择该方法。在实际应用中，可以首先选择剔除缺失值较多的样本，再采用其他方式填补剩余的缺失值。

在本实验数据中，还有缺失项的样本共有 16 个，剔除后，剩余 184 个。输出结果在..\result\Analysis_Delete.csv 文件中，可视化结果在..\result\picture\Delete 中。

(2) 用最高频率值来填补缺失值

采用众数填补缺失值，优点简单速度快，但可能存在较大的数据偏差。

最高频率填充输出结果在..\result\Analysis_Frequency.csv 文件中，可视化结果在..\result\picture\Frequency 中。

(3) 通过属性的相关关系来填补缺失值

在本实验中，200 个样本值中，16 个样本有缺失项，8 个数值属性都有缺失值。观察数据后发现样本 62 和 199 含有太多的缺失值，为不具备参考的样本，剔除后，NH4 和 NO3 就没有缺失值了。剔除后有缺失值的属性为 mxPH、mnO2、Cl、PO4、Chla，属性 NH4、NO3、oPO4 无确实值。通过探寻变量之间的相关关系，使用变量的相关关系填补缺失值。在本实验计算所得变量之间的相关值矩阵结果保存在..\result\CorrelationMatrix.txt 中，如下图:

```
      mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH 1
mnO2    1
Cl         1
NO3           1
NH4           , 1
oPO4    .  .        1
PO4     .  .        * 1
Chla .                  1
a1         .        . .    1
a2   .                    .     1
a3                                1
a4      .           . .             1
a5                                     1
a6            .  .                     .  1
a7                                           1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

图 7 变量之间的相关值矩阵

通过变量之间的相关值矩阵，发现变量 PO4 和 oPO4 的相关性>0.9，非常相关；NH4 和 NO3 的相关性>0.6，比较相关(剔除 62、199 后不存在缺失项，不用填补)；oPO4 和 mnO2、oPO4 和 Cl、Chla 和 mxPH 的相关性>0.3，一般相关。本实验中相关系数的计算结果保存在..\result\CorrelationCoefficient.txt 中。

首先利用 PO4 和 oPO4 的相关性，填补 PO4 的缺失值。经过计算，PO4 和 oPO4 的线性相关公式为：PO4=42.897+1.293*oPO4。

还有 mxPH、mnO2、Cl、Chla 含有缺失值，可以通过探索数值属性与标称变量之间的关系，来填补缺失值。这种分析方法比较繁琐，但是可以应用到少量名义变量的较小数据集的分析中。如 mnO2 的缺失项为 season=spring，size=small，speed=high，通过图示观察，该条件下(图中 3 行 2 格)数据主要集中在 8-10 之间，靠近均值 Mean=9.118，所以采用均值填补该缺失项。但是 mxPH 在这三个条件下的变化不明显。季节、河流大小和速度引起的 mnO2、mxPH 的变化，如下图：

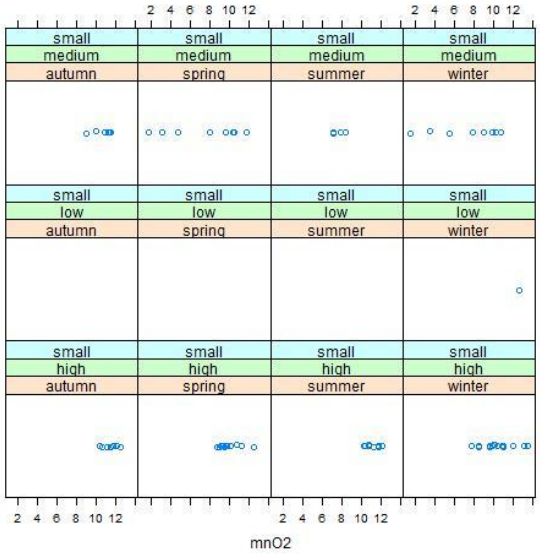

图 8 变量 mnO2 在季节、河流大小和速度条件下的分布

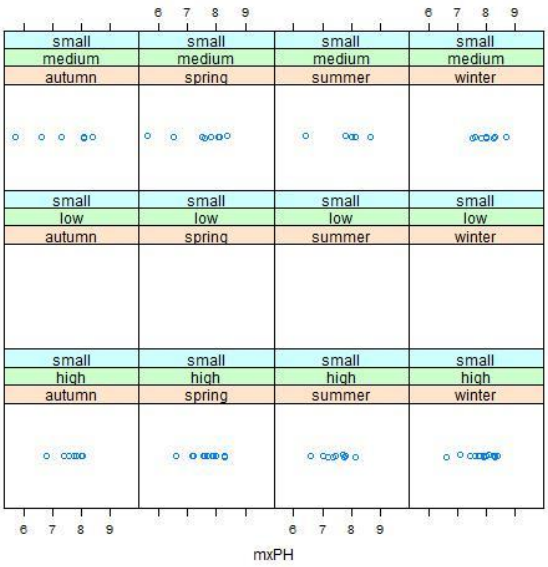

图 9 变量 mxPH 在季节、河流大小和速度条件下的分布

同时由于Cl、Chla的缺失项较多，观察结果不明显。本实验仍采用mnO2和oPO4、Cl和oPO4、Chla和mxPH的线性相关性来填补缺失项，由于相关性较小，会存在偏差，这样做的目的是为了对比填充后的结果。经过计算，Chla和mxPH的线性相关公式为:Chla=-139.4+19*mxPH，mxPH=7.92896+0.01047*Chla。mnO2和oPO4的线性相关公式为:mnO2=9.93341-0.01093*oPO4。Cl和oPO4的线性相关公式:Cl= 28.4771+ 0.1992*oPO4。

相关性填充输出结果在..\result\Analysis_Correlation.csv 文件中，可视化结果在..\result\picture\Correlation中。

(4) 通过数据对象之间的相似性来填补缺失值

相似性经常由描述观察值的多远度量空间的变量所定义，在本实验中，采用最常用的欧式距离，其公式定义为$d(\mathrm{x},\mathrm{y})=\sqrt{\sum_{i=1}^{P}(\mathrm{x}_i-\mathrm{y}_i)^2}$ 。

采用欧氏距离寻找与任何含有确实值的样本最相似的10个样本，并用它们来填补缺失值。采用这些最相似数据的加权平均，权重的大小随着距待填补缺失值的样本的距离增大而减小。本实验中，采用高斯核函数从距离获得权重，设距离为d，则其权重为$w(\mathrm{d})=\mathrm{e}^{-d}$。

相似性填充输出结果在..\result\Analysis_Similarity.csv 文件中，可视化结果在..\result\picture\Similarity中。