作业三、缺陷数据集分析

MG1733018 郭肇强

1 任务描述

在软件开发时会产生很多缺陷，软件缺陷预测是一种行之有效的软件质量控制手段，通过对缺陷数据集的分析，可以提供有用的缺陷预测指标进行缺陷预测。本实验是通过编写 R 脚本分析 wmc，dit，noc，cbo，rfc 和 lcom 这 6 种度量值的缺陷预测能力。使用的数据集为 xalan2.4：https://zenodo.org/record/268436/files/xalan-2.4.csv。

2 实验内容

2.1 收集描述性统计信息最小值、25%处值、中位值、75%处值、最大值、平均值、偏度(skewness)和峰度(kurtosis)

（1）统计信息计算方法

平均值：

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.1}$$

中位值：

$$Median = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & ,n \in odd \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right] & ,n \in even \end{cases} \tag{3.2}$$

方差：

$$S^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.3}$$

偏度：

$$skewness = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{(n-1)S^3} \tag{3.4}$$

峰度：

$$kurtosis = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{(n-1)S^4} - 3 \tag{3.5}$$

（2）实现方法：calc_statis(features) features 为 6 类度量指标

```
calc_statis(features)                    #a.描述性统计信息,输出结果"statis.csv"
```

（3）输出结果

表 1　描述性统计信息

name	min	Q1.	median	Q3.	max	mean	skewness	kurtosis
wmc	0	3	6	12.5	123	11.44952	3.478202	15.08622
dit	1	1	2	4	8	2.565698	0.656867	-0.29915
noc	0	0	0	0	29	0.608575	7.332323	63.16812
cbo	0	4	8	18	171	14.49793	3.472056	16.41106
rfc	0	8	19	41	355	30.16183	3.014723	14.6192
lcom	0	0	3	22.5	6589	130.0816	7.684377	67.13383

2. 收集度量数据与 bug 数据的相关系数及其显著性（Spearman、Pearson）

（1）**Spearman 相关系数** 在统计学中，斯皮尔曼等级相关系数用来估计两个变量 X、Y 之间的相关性，其中变量间的相关性可以使用单调函数来描述。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数时（即两个变量的变化趋势相同），两个变量之间的 ρ 可以达到+1 或-1。计算方法如（3.6）。

斯皮尔曼等级相关系数对数据条件的要求没有皮尔逊相关系数严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关系数来进行研究。

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{3.6}$$

（2）**Pearson 相关系数** 皮尔逊相关也称为积差相关（或积矩相关）是英国统计学家皮尔逊于 20 世纪提出的一种计算直线相关的方法。假设有两个变量 X、Y，那么两变量间的皮尔逊相关系数可通过以下公式计算(其中 E 是数学期望，cov 表示协方差)：

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{3.7}$$

（3）**t 统计检验**

$$t = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} \tag{3.8}$$

（4）实现方法：calc_cor(features) features 为度量数据

```
calc_cor(features)                    #b.计算相关系数及显著性统计,输出结果"cor.csv"
```

（5）输出结果 （r 为相关系数， T 为显著性结果）

表 2　相关系数及显著性水平

r.spearman	T.spearman	r.pearson	T.pearson
0.314245034	0.662027097	0.37879231	0.818584115
-0.026123016	-0.052263867	-0.00186037	-0.00372075
0.090944259	0.182645404	0.054916	0.109997988
0.217623951	0.445935783	0.22354421	0.458696293
0.356341649	0.76275371	0.45929364	1.034113835
0.259251617	0.536858525	0.30757572	0.646490996

3.4. 使用 10 种机器学习方法建立多变量的缺陷预测模型，利用 10x10 交叉验证评价模型的分类性能（AUC）和排序性能（CE）。

```
learners <- c(
    "classif.naiveBayes",      #朴素贝叶斯分类器
    "classif.svm",             #支持向量机
    "classif.gbm",             #梯度推进机
    "classif.lda",             #线性判别分析
    "classif.mlp",             #多层感知器
    "classif.randomForest",    #随机森林
    "classif.rpart",           #决策树
    "classif.glmnet",          #GLM with Lasso or Elasticnet Regularization
    "classif.nnet",            #神经网络
    "classif.multinom"         #多元回归
    )
```

图 1　10 种机器学习方法

（1）分类性能 AUC

AUC 可通过对 ROC 曲线下各部分的面积求和而得。假定 ROC 曲线是由坐标为$\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$的点按序连接而形成$(x_1=0,x_m=1)$，则 AUC 可估算为

$$AUC=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1}-x_i)\cdot(y_i+y_{i+1}) \tag{3.9}$$

（2）排序性能 CE

其中$Area_\pi(Random)$是 0.5，因此 CE 和 AUC 之间是正相关。

$$CE_\pi(model)=\frac{Area_\pi(model)-Area_\pi(Random)}{Area_\pi(optimal)-Area_\pi(Random)} \tag{3.10}$$

（3）实现方法:make_and_evaluate_model()

```
make_and_evaluate_model()          #c.构建模型并评估
```

（4）输出结果

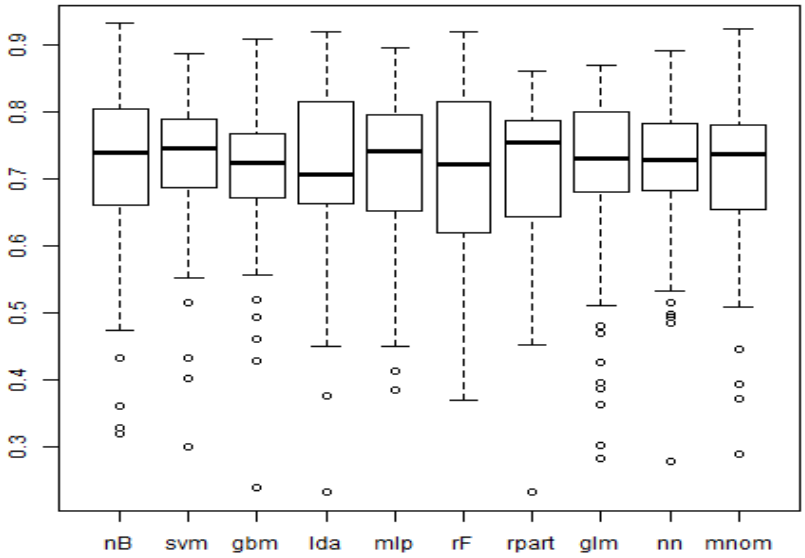

图 2　算法性能

5. 利用 CD 图比较这 10 种模型在统计上的差别（plotCD）

（1）实现方法：graph_cd(graph_data)

```
graph_cd(graph_data)                              #e.作CD图
```

（2）输出结果

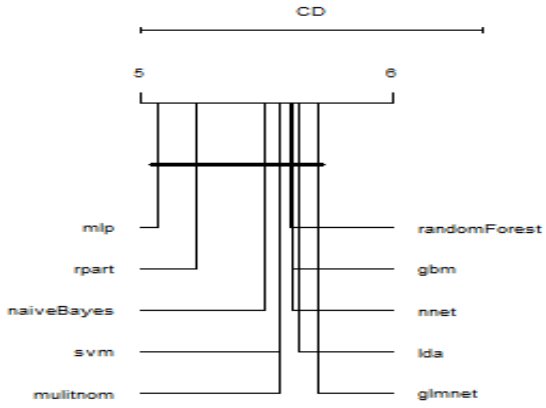

图 3 CD 图

6. 利用 Algorithm 图比较这 10 种模型在统计上的差别

（1）实现方法：graph_algorithm(graph_data)

```
graph_algorithm(graph_data)                    #f.作Algorithm图
```

(2) 输出结果

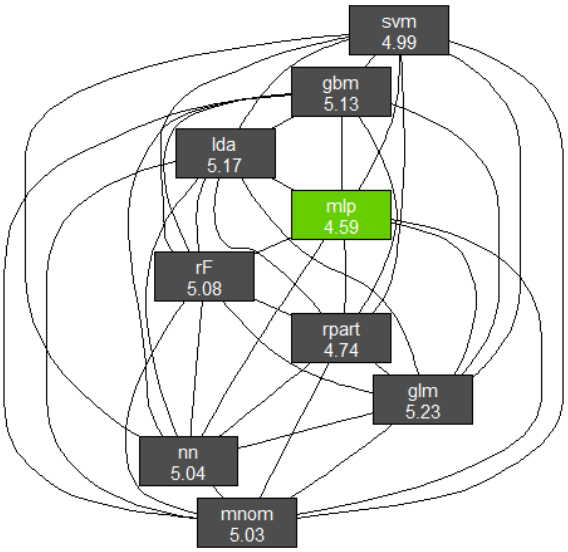

图 4　算法图

7. 利用 heatmap 展示 10 个模型在 100 个测试集上的结果（行为模型，列为结果）

```
graph_heatmap(graph_data)                      #g.作heatmap图
```

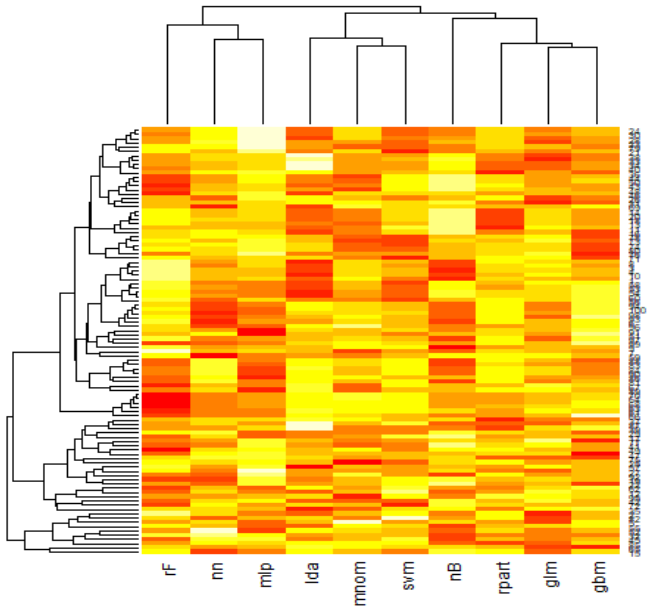

图 5　热力图

参考文献资料

[1] An Introduction to R
[2] https://mlr-org.github.io/mlr-tutorial/devel/html/index.html

[3] Statistical Comparison of Multiple Algorithms in Multiple Problems

[4] 赵东晓 周毓明.无监督缺陷模块序列预测模型：一个工作量感知的评价[J]中国科技论文在线

[5] 周志华.机器学习[M]. 清华大学出版社