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：

中位值：

方差：

偏度：

峰度：

（2）实现方法：calc\_statis(features) features为6类度量指标



（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。

（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)：

（3）**t 统计检验**

（4）实现方法：calc\_cor(features) features为度量数据



（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）。

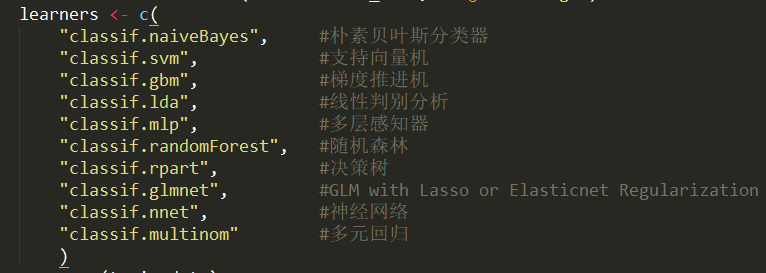


图1 10种机器学习方法

（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曲线是由坐标为的点按序连接而形成，则AUC可估算为

（2）排序性能CE

其中是0.5，因此CE和AUC之间是正相关。

（3）实现方法:make\_and\_evaluate\_model()



（4）输出结果

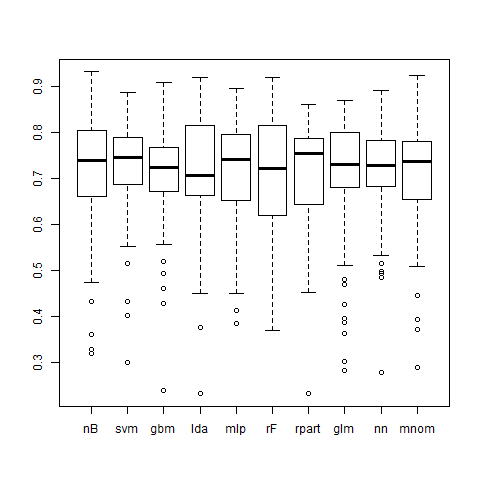


图2 算法性能

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)



（2）输出结果

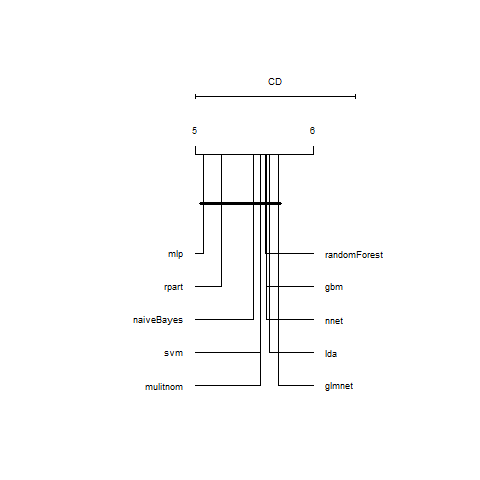


图3 CD图

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)



(2) 输出结果

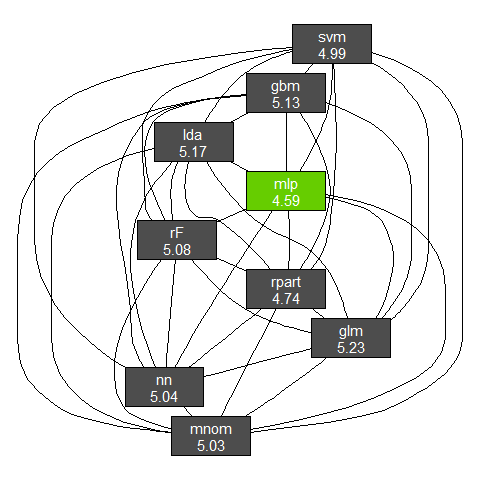


图4 算法图

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）



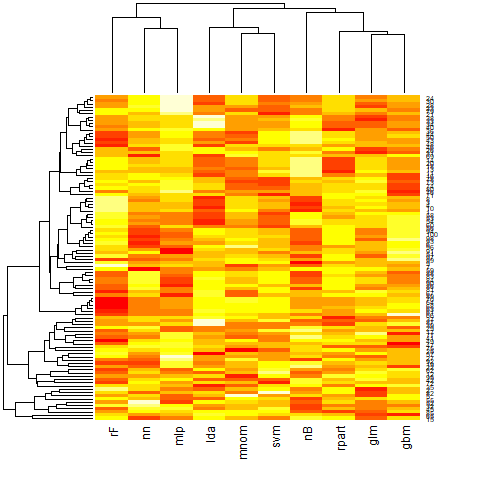


图5 热力图

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