缺陷数据集分析

任务描述

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。

实验内容

**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)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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）



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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.使用10种机器学习方法建立多变量的缺陷预测模型（不进行特征选择）

4.利用10x10交叉验证方法评价模型的分类性能（AUC）和排序性能（CE）

5.利用CD图比较这10种模型在统计上的差别（plotCD）

6.利用Algorithm图比较这10种模型在统计上的差别

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）