4 steps for analyzing the data

把六个metric的数据收集后，需要对数据进行相关性分析。相关性分析采用皮尔森相关系数以及斯皮尔曼相关性系数[如果在introduction上没有介绍的话需要写上两者的公式(presentation的ppt上有)以及参考文献]

数据分析的步骤如下：

1.确定哪两个metric进行相关性的比较分析，确定两个metric是数据哪个层级的数据（例如，package level，class level）。从收集到的数据中提取出特定项目的metric数据。

2.把收集到的metric数据导入进行皮尔森相关性系数以及斯皮尔曼相关性系数的相关性分析的Matlab程序中，得到相关系数，并生成数据点的分布图。

3.对五个项目的特定metric的相关系数的结果进行比较，得出最具普遍性的结论。

5 结果分析

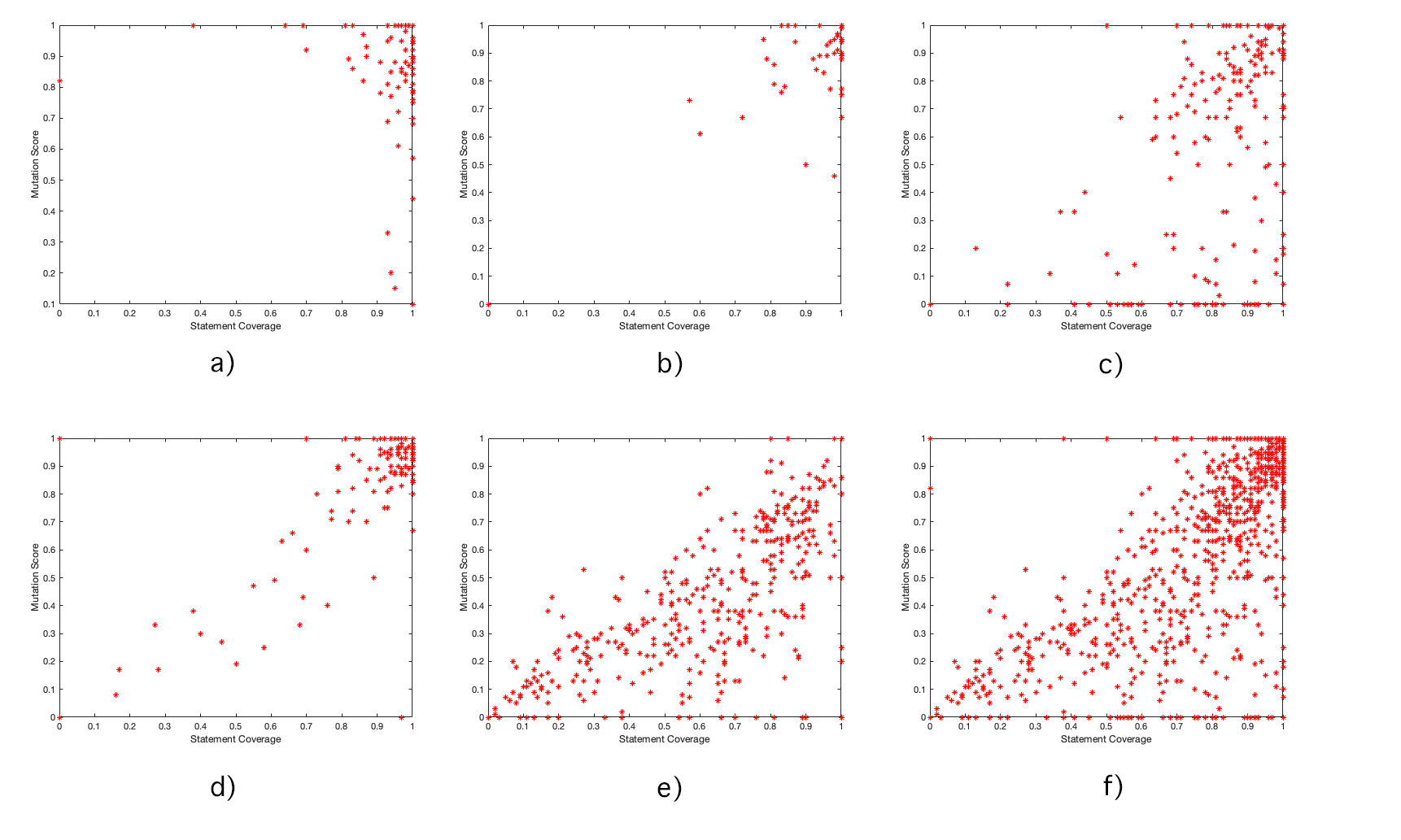
5.1 Correlation between Metric 1&2 and Metric 3

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | R(pearson) of metric 1&3 |
| Total 5 project | 1063 | 0.7476 |
| Apache commons Lang | 89 | -0.0564 |
| Apache commons codec | 52 | 0.8027 |
| Apache commons collections | 264 | 0.4510 |
| Apache commons configuration | 177 | 0.8266 |
| JFreeChart | 481 | 0.7996 |
| Apache commons Lang | 11 sets PackageLevel data | 0.3152 |

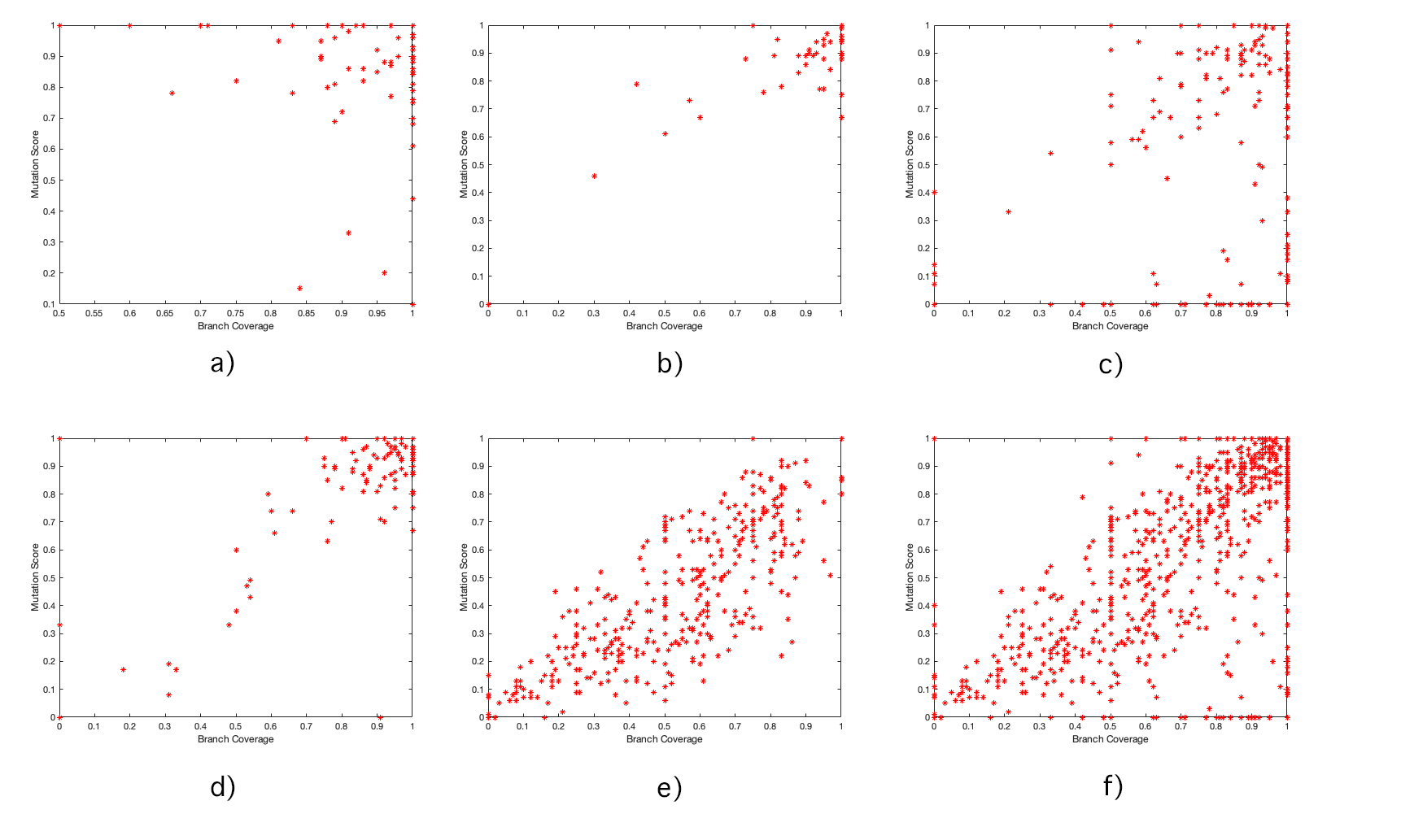
表1 metric1和metric3的皮尔森相关系数

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | R(pearson) of metric 2&3 |
| Total 5 project | 899 | 0.7707 |
| Apache commons Lang | 75 | -0.0847 |
| Apache commons codec | 47 | 0.8674 |
| Apache commons collections | 206 | 0.3714 |
| Apache commons configuration | 143 | 0.753 |
| JFreeChart | 428 | 0.7996 |
| Apache commons Lang | 11 sets PackageLevel data | 0.8627 |

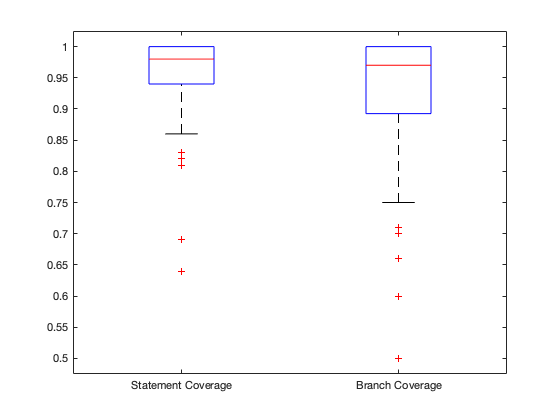
表2 metric2和metric3的皮尔森相关系数



**Figure1** metric1和metric3的类级别的数据分布图a) Apache commons Lang b) Apache commons codec c) Apache commons collections d) Apache commons configuration e) JFreeChart f）Total five project class level data



**Figure2** metric2和metric3的类级别的数据分布图a) Apache commons Lang b) Apache commons codec c) Apache commons collections d) Apache commons configuration e) JFreeChart f）Total five project class level data



**Figure3** Apache commons Lang metric1&2 boxplot

对metric1 和 metric3 以及metric2和metric3的Class level数据进行相关性分析，五个项目的数据点分布的figure如figure1 和figure2所示，通过figure1和figure2以及以上的表1和表2可以看出，除了Project Apache commons Lang的相关性系数和其他四个project的相关性系数有明显的区别外，其他四组的相关性都是比较strong而且呈现相关性的方向是正相关。

对此，我们对五个项目的class level的数据汇总进行总的分析，得出的*R(pearson)* of metric 1&3 of Total 5 project is 0.7476, *R(pearson)* of metric 2&3 of Total 5 project is 0.7707。

对项目Apache commons Lang的为何相关系数很小进行深入的分析，在Apache commons Lang的Metric1和metric2的大小集中在90%以上,如figure3所示，数据分布过于集中，很难形成良好的相关性比较，很容易导致相关系数在Apache commons Lang上出现偏差。对此，我们在表1表2中特别列出了Apache commons Lang的package level的11组数据生成的metric1&2和metric3的相关性系数较大而且正相关，结合四组项目的大小相似的相关系数以及五组数据的普遍性，可以认为Apache commons Lang的class level上相关系数属于异常数据，不具有普遍性，可以忽略。

因此，由以上五组项目皮尔森相关性系数可以看出，相关性是很强的正相关。

5.2Correlation between Metric 1&2 and Metric4

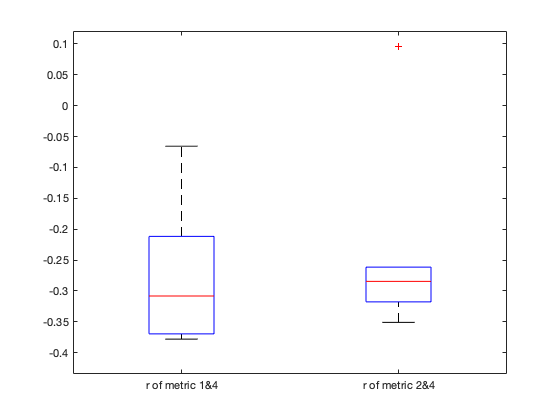
对于metric1&2 和 metric 4的相关性分析，我们采用斯皮尔曼相关系数rs进行分析。

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | rs of metric 1&4 |
| Total 5 project | 1663 | -0.3556 |
| Apache commons Lang | 246 | -0.2116 |
| Apache commons codec | 89 | -0.2605 |
| Apache commons collections | 474 | -0.3780 |
| Apache commons configuration | 306 | -0.3694 |
| JFreeChart | 548 | -0.0655 |

表3 metric1和metric4的斯皮尔曼相关系数

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | rs of metric 2&4 |
| Total 5 project | 1174 | -0.2705 |
| Apache commons Lang | 162 | -0.2985 |
| Apache commons codec | 59 | -0.3509 |
| Apache commons collections | 319 | -0.2614 |
| Apache commons configuration | 197 | -0.3177 |
| JFreeChart | 437 | 0.0958 |

表4 metric2和metric4的斯皮尔曼相关系数

****

**Figure4** boxplot of rs of metric1&4 and metric 2&4

由以上表3和表4的五个项目的metric1和metric4,metirc2和metric4的斯皮尔曼相关系数以及figure4可以看出，大部分项目的metric1和metric4以及metric2和metric4的斯皮尔曼相关系数处于-0.3附近。我们可以由这两个表格的斯皮尔曼相关系数知道，The correlation between metric 1&2 and 4 is negative and the strength of the association is good but not very strong.

5.3 Correlation between Metric 1&2 and Metric 6

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | R(pearson) of metric 1&6 |
| Apache commons Lang | 126 | -0.0544 |
| Apache commons codec | 60 | -0.0761 |
| Apache commons collections | 270 | -0.0237 |
| Apache commons configuration | 186 | 0.0404 |
| JFreeChart | 524 | 0.0328 |

表5 metric1和metric6的皮尔森相关系数

|  |  |  |
| --- | --- | --- |
| Project | Sets of data(Class level) | R(pearson) of metric 2&6 |
| Apache commons Lang | 108 | 0.0541 |
| Apache commons codec | 47 | -0.0734 |
| Apache commons collections | 206 | -0.0245 |
| Apache commons configuration | 143 | -0.0031 |
| JFreeChart | 428 | 0.0694 |

表6 metric2和metric6的皮尔森相关系数

Metric1&2与metric6的皮尔森相关系数在表5和表6上，五个项目的的metric1&2和metric6的皮尔森相关系数的绝对值均小于0.01。因此，我们推断metric6和metric1&2几乎没有相关性。

5.4 Correlation between Metric 5 and Metric 6

|  |  |  |
| --- | --- | --- |
| Project(Version-Verison) | Metric5 BMI | Metric 6 Change proneness |
| Apache commons Lang 3.0-3.6 | 10.833 | 0.00591716 |
| Apache commons Lang 3.6-3.7 | 43.333 | 0.020833333 |
| Apache commons Lang 3.7-3.8 | 33.9683 | 0.017241379 |
| Apache commons codec 1.10-1.11 | 30.5556 | 0.03125 |
| Apache commons codec 1.11-1.12 | 44.4444 | 0.041666667 |
| Apache commons codec 1.9-1.10 | 100 | 0.025641026 |
| Apache commons collections 3.2-4.0 | 40.3175 | 0.00177305 |
| Apache commons collections 4.0-4.1 | 38.611 | 0.005076142 |
| Apache commons collections 4.1-4.3 | 41.6667 | 0.003030303 |
| Apache commons configuration 2.1-2.2 | 66.667 | 0.00990099 |
| Apache commons configuration 2.2-2.3 | 15.7576 | 0.071428571 |
| Apache commons configuration 2.3-2.4 | 3.0303 | 0.005235602 |
| Jfreechart 1.0.18-1.0.19 | 250 | 0.045454545 |
| Jfreechart 0.19-1.5.0 | 66.667 | 0.000770416 |

表7 收集到的五个项目不同版本metric5和metric6的数据

对以上的14组数据进行皮尔森相关系数的计算，R（Pearson）的值为0.2732，可以认为metric5和metric6具有中等强度的正相关性。

5.5相关性分析结论

通过5.1我们知道The Correlation between Metric 1&2 and Metric 3 is positive and very strong.

We can conclude that suites with higher Statement or Branch coverage can show high mutation score. This conclusion is consistent with the rationale that test suites with higher coverage can show better test suite effectiveness.

通过5.2我们知道The correlation between metric 1&2 and 4 is negative and the strength of the association is good but not very strong. We can conclude that classes with higher Cyclomatic Complexity show lower Statement/Branch coverage . This conclusion is consistent with the rationale that classes with higher complexity are less likely to have high coverage test suites.

通过5.3我们知道Metric1&2和metric6的皮尔森相关系数非常小，绝对值不大于0.1。因此我们认为Metric1&2和metric6几乎没有相关性。我们认为，每个类的Statement/Branch coverage和change proneness无关。

通过5.4我们知道Metric5和metric6的皮尔森相关系数呈现正相关，而且相关强度中等。

We conclude that on the project’s level, project with higher Backlog Management Index might show higher **change proneness.**