4. Steps for analyzing the data

It is necessary to do correlation analysis after collecting the data for these 6 metrics. The correlation analysis was adopted by the Pearson correlation coefficient and Spearman correlation coefficient. (If there is no introduction in the ‘Introduction’ part, the formulas and references for both should be included)

The Steps of data analysis are as follows:

(1) Determining which two metrics are used for correlation comparative analysis, and determining which level of data they are ( such as, package level, class level ). Extracting the metric data of specific project from the collected data.

(2) Importing the collected metric data into ‘Matlab’ for both Pearson correlation coefficient analysis and Spearman correlation coefficient analysis. Collecting the correlation coefficient and generating the distribution map of the data points.

(3) Comparing the results of the specific metric correlation coefficients of the five projects and drawing the most general conclusion.

5.Results analysis

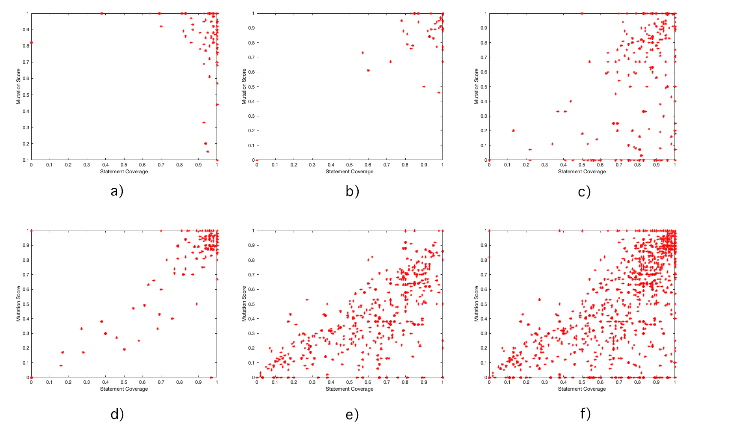
5.1 Correlation between Metric 1 & 2 & 3.

|  |  |  |
| --- | --- | --- |
| 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 Package-Level data | 0.3152 |

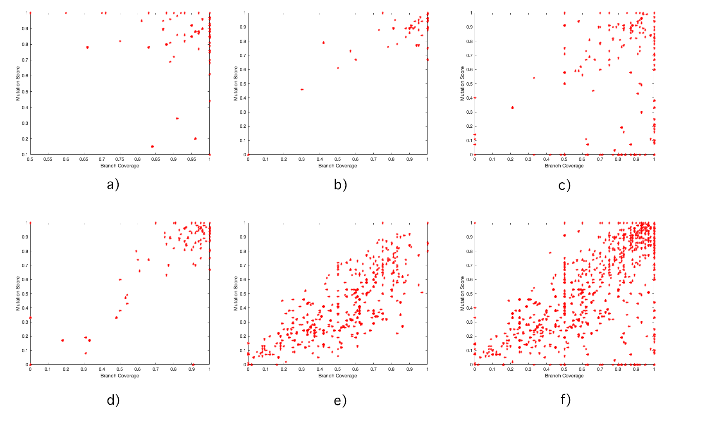
Table1. The Pearson correlation coefficient between Metric 1 & 3

|  |  |  |
| --- | --- | --- |
| 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 Package-Level data | 0.8627 |

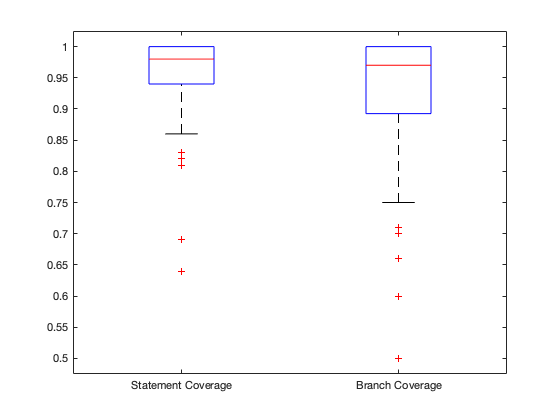
Table2. The Pearson correlation coefficient between Metric 2 & 3



**Figure1** Data distribution diagram of Class-Level between Metric 1 & 3 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** Data distribution diagram of Class-Level between Metric 2 & 3 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 metric 1 & 2 boxplot

Correlation analysis of Class level data for metric1 and c3 and for c2 and c3, the distribution of data points for the five projects is shown in the figures (**Figure1,Figure2,Figure3**).

It can be seen from figure 1 and 2, as well as table 1 and table 2 above that the correlation of the four groups is relatively strong and the direction of the correlation is positive except the ‘Apache Commons Lang’ project. In fact, the correlation coefficient of ‘Apache Commons Lang’ project is obviously different from the other four projects.

Hence, we summarized and analyzed the class-level data of the five projects, and concluded that R(Pearson) of metric 1&3 of total 5 projects is 0.7476,

and R(Pearson) of metric 1^3 of total 5 projects is 0.7707.

There is an in-depth analysis which is made on why the correlation coefficient of the project Apache Commons Lang is very small. The size of the Apache Commons Lang Metric 1 and Metric 2 is concentrated at more than 90%, as shown in figure3. In fact, the data distribution is too centralized to form a good correlation comparison, and it is easy to cause the deviation of the correlation coefficient on Apache Commons Lang. So the figure1 and figure 2 shows that the correlation coefficient is large and positive for metric1&2 and metric 3 based on the 11 sets data of package-level in Apache Commons Lang. After combining with the size-similar correlation coefficients of the four groups of projects and the universality of the five groups of data, it can be seen considered that the correlation coefficient on the class-level of Apache Commons Lang belongs to abnormal data, which is not universal and can be ignored. Therefore, it can be seen from the Pearson correlation coefficient of the above five groups of projects that the correlation is very strong and positive.

5.2 Correlation between Metric 1&2 and Metric4.

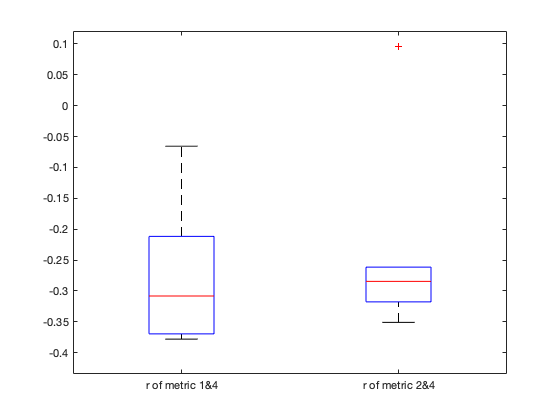
The correlation analysis of metric 1&2 and metric 4 was carried out using the Spearman correlation coefficient rs.

|  |  |  |
| --- | --- | --- |
| 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 |

Table3. Spearman correlation coefficients for metric 1 and metric 4

|  |  |  |
| --- | --- | --- |
| 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 |

Table4. Spearman correlation coefficients for Metric 2 and Metric 4

****

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

As can be seen from figure4 and the Spearman correlation coefficients of metric 1&4, 2&4 of the five items in tables3 and table4 above, the Spearman correlation coefficients of most items are around -0.3. So we can know from these two tables that 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 |

Table5. Pearson correlation coefficients for Metric 1 and Metric 6

|  |  |  |
| --- | --- | --- |
| 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 |

Table6. Pearson correlation coefficients for Metric 2 and Metric 6

The Pearson correlation coefficients for Metric 1&2 and Metric 6 are shown in table 5 and table 6. The absolute Pearson correlation coefficients for all five items are less than 0.01. Consequently, we inferred that Metric 6 and Metric 1&2 had little correlation.

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 |

Table7. Collecting Metric 5 and Metric 6 data for 5 different versions of the project

The Pearson correlation coefficient was calculated for the above 14 sets of data, and the value of R(Pearson) was 0.2732, so it shows that the positive correlation between Metric 5 and Metric 6 was moderately strong.

5.5 Conclusions of correlation analysis

According to ‘5.1’, it shows that 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.

According to ‘5.2’, it depicts that 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.

According to ‘5.3’, it describes that the Pearson correlation coefficients for Metric 1&2 and Metric 6 were very small, even not greater than 0.1 in absolute value. Therefore, we consider that Metric 1&2 and Metric 6 are almost uncorrelated. We think that the Statement/Branch coverage of each class has nothing to do with change proneness.

According to ‘5.4’, it depicts that the Pearson correlation coefficients of the Metric 5 and Metric 6 were positively correlated and moderately strong. We conclude that on the project-level, project with higher Backlog Management Index might show higher change proneness.