Finding Branch Correlations

Jaivardhan Kapoor (150300) Shivansh Rai (14658)

2017/09/12

1 Problem Statement

A running problem in program analysis is to compute branch similarity of control nodes, e.g. if statements. Since data-driven methods are to be used, the aim is to find branch correlations using ouptut from the test data applied to the test program.

2 Methodology

We will be using ROSE, compiler framework that is capable of changing the source code of a program by modifying its intermediate representation(IR). The process will consist of the following steps:

- Modifying the source code of the test program using ROSE. We wil be adding multiple helper/utility functions for the bookkeeping and evaluation of the if-else statement branch hits, and will modify the output of the program to output codes corresponding to the type of hits of each branch. Since the program will have outputs of its own, we will consider only the last line of the output, and extract it using shell scripting.
- Converting the output of the modified test program to a csv file in the form of a contingency table. The rows of the table will be the line numbers of the if statements, and the number of occurences of each type of hit will be depicted as columns. The different types of behaviours of the if statements are:
  - 0 - The branch never hits
  - 1 - The branch hits and is always true
  - 2 - The branch hits and is always false
  - 3 - The branch hits and can be either true or false

  We will be using the generated contingency table to find out thee correlations between branches using *Pearson's Chi-square(χ^2) test*. The signifiance of the p-value will be set to 0.05, and the required p-values will be calculated using R. Please refer to the README file in the folder for instructions on running the code.

3 Hypothesis Testing

The contingency table is given below:

Branches	Hit 0	Hit 1	Hit 2	Hit 3
Line 84	0	498	0	0
Line 106	216	158	124	0
Line 143	216	158	124	0
Line 148	0	282	216	0
Line 157	216	0	282	0
Line 163	216	58	224	0
Line 184	274	37	187	0

We observe that the there are repeating values in many of the table cells, which suggests that the branches are highly correlated.
On appying Pearson's χ^2 test for independence, we obtain a matrix of p-values for each pair of line numbers(or branches). The level of significance is taken as 5%.
For hypothesis testing, we assume that the null hypothesis H_0 is that each pair of considered branches are independent. If the null hypothesis fails to be rejected, we will accept the alternative hypothesis H_1 that the branches are correlated.
Our observation is that the p-value matrix contains no value that is lesser than the level of significance(5%).

Therefore, we are led to conclude that none of the branches are independent of each other, and they are strongly correlated as evident from the correlation matrix obtained from the R program.