T-Wise Test Generation and Prioritization For Large Software Product Line: Technical Report

Christopher Henard and Gilles Perrouin

August 18, 2012

Abstract

This document presents the distance and fitness functions used in the paper entitled "Bypassing the Combinatorial Explosion: Using Similarity to Generate and Prioritize T-wise Test Suites for Large Software Product Lines". In particular, it is shown how these functions are useful in terms of t-wise coverage.

Contents

1 Representation of the products of a feature model **1**

2 A distance between two prodcuts **1**

2.1 Evaluating the t-wise coverage of two products 2

3 A fitness function for N products **2**

3.1 Comparing two N-uplets of products in terms of t-wise coverage 3

1 Representation of the products of a feature model

One product or one configuration is represented as a set of n features of a feature model as $P = \{\pm f_1, ..., \pm f_n\}$, where $+f_i$ indicates a feature which is selected by this product, and $-f_i$ an unselected one. Table 1 illustrates an example of three products and four features.

For instance, product $P_1 = \{+f_1, +f_2, +f_3, -f_4\}$ supports all the features except f_4.

		Features			
		f_1	f_2	f_3	f_4
Products	P_1	×	×	×	
	P_2	×	×		×
	P_3	×		×	

Table 1: Example of Three Products For an FM of Four Features

2 A distance between two prodcuts

We use the Jaccard distance [1] d between two products of n features:

$$d : \begin{array}{ccc} P \times P & \longrightarrow & [0,1] \\ (P_i, P_j) & \longmapsto & 1 - \dfrac{\#P_i \cap P_j}{\#P_i \cup P_j}, \end{array}$$

where $P_i, P_j \in P$, where n denotes the number of features of the feature model, where $\#A$ denotes the cardinality of the set A and with $\#P_i \cup P_j > 0$.

For instance, with reference to Table 1, $n = 4$ and $P_1 = \{+f_1, +f_2, +f_3, -f_4\}$, $P_2 = \{+f_1, +f_2, -f_3, +f_4\}$ and $P_3 = \{+f_1, -f_2, +f_3, -f_4\}$. Thus:

- $d(P_1, P_2) = 1 - \frac{\#\{+f_1,+f_2\}}{\#\{+f_1,+f_2,+f_3,-f_3,+f_4,-f_4\}} = 1 - \frac{2}{6} \approx 0.67$,
- $d(P_1, P_3) = 0.4$,

- $d(P_2, P_3) = \approx 0.86$.

In this example, P_1 and P_3 are the most similar products (i.e. they share the lowest distance), whereas P_2 and P_3 are the most dissimilar ones.

Besides, the distance d is a normalization between 0 and 1 of the following distance d':

$$d' : \begin{array}{ccc} P \times P & \longrightarrow & [0, \#P_i \cup P_j] \\ (P_i, P_j) & \longmapsto & \#P_i \cup P_j - \#P_i \cap P_j. \end{array}$$

In our context, the exact distances values are not important. What is relevant for us is the order between two distances, meaning that if we have N products and k Jaccard distances, we have $d^1(P_1, P_2) < ... < d^k(P_{N-1}, P_N) \Leftrightarrow d'^1(P_1, P_2) < ... < d'^k(P_{N-1}, P_N)$ because the normalization does not affect the ordering between the distances. In a similar way, $d^1(P_1, P_2) > ... > d^k(P_{N-1}, P_N) \Leftrightarrow d'^1(P_1, P_2) > ... > d'^k(P_{N-1}, P_N)$. In the following, we'll use d' to simplify the demonstration.

2.1 Evaluating the t-wise coverage of two products

The t-wise coverage expresses the ability of the products to cover the t-sets of features of the feature model. A t-set is a valid combination of t (selected or unselected) features of the feature model. This coverage can be evaluated as the number of t-set covered by the products.

Each feature model of n features has at most $\binom{2n}{t}$ valid t-sets. Each product P covers exactly $C(P, n, t) = C(P, \#P, t) = \binom{n}{t}$ t-sets. As a result, the t-wise coverage of two products P_i and P_j is defined by the number of t-sets covered by these two products minus the common t-sets covered by these products:

$$\begin{aligned} Coverage(P_i, P_j) &= C(P_i, \#P_i, t) + C(P_j, \#P_j, t) - C(P_i \cap P_j, \#P_i \cap P_j, t) \\ &= C(P_i, \#P_i, t) + C(P_j, \#P_j, t) - C(P_i \cap P_j, \underbrace{\#P_i \cup P_j - d'(P_i, P_j)}_{x}, t). \end{aligned}$$

Proposition 1. *The distance d' (and thus d) is a good indicator of the t-wise coverage between two products, i.e. $Coverage(P_i, P_j)$ increases with $d'(P_i, P_j)$.*

Proof. We have, in particular:

- When $d'(P_i, P_j) = 0$, then $x = n$ and $\#P_i \cap P_j = n$. Thus, $P_i = P_j$ and the coverage is at its minimum value: $Coverage(P_i, P_j) = C(P_i, n, t) = C(P_j, n, t) = \binom{n}{t}$ t-sets.
- When $d'(P_i, P_j) = n$, then $x = n$ and $\#P_i \cap P_j = 0$. Thus, the coverage is at its maximum value: $Coverage(P_i, P_j) = C(P_i, n, t) + C(P_j, n, t) = 2\binom{n}{t}$ t-sets.

Thus:

- $Coverage(P_i, P_j) \geq Coverage(P'_i, P'_j) \Leftrightarrow d'(P_i, P_j) \geq d'(P'_i, P'_j)$.
- $Coverage(P_i, P_j) \leq Coverage(P'_i, P'_j) \Leftrightarrow d'(P_i, P_j) \leq d'(P'_i, P'_j)$.

□

As a result, the above-mentioned distance d' (and thus d) is relevant to evaluate the coverage of two products without computing the t-sets.

3 A fitness function for N products

We define a fitness function f based on d to evaluate the quality of N products $P_1, ..., P_N$:

$$f : \begin{array}{ccc} P^N & \longrightarrow & \mathbb{R}_+ \\ (P_1, ..., P_N) & \longmapsto & \sum_{j>i\geq 1}^{N} d(P_i, P_j). \end{array}$$

3.1 Comparing two N-uplets of products in terms of t-wise coverage

For a given and fixed value of N, this fitness function allows to choose between two N-uplets products $S_1 = (P_1, ..., P_N)$ and $S_2 = (P'_1, ..., P'_N)$ which set potentially has the highest t-sets coverage. To this end, we evaluate $f(S_1)$ and $f(S_2)$ and we take the set where f is maximum.

Proposition 2. *The fitness function f is a good indicator of the t-wise coverage between $N \geq 2$ products, i.e. $Coverage(P_1, ...P_N)$ generally increases with $f(P_1, ..., P_N)$.*

Proof. The fitness function d corresponds to the sum of the distances pairwise between the products:

$$f(P_1, ..., P_N) = d(P_1, P_2) + d(P_1, P_3) + ... + d(P_{N-1}, P_N).$$

In other words, f corresponds to the evaluation of d between each couple of products. Since it's a sum and given the properties of d, it is direct that the value of f will be maximum when all the terms of the sum are maximum, i.e. when the distance d among any two products is maximum.

□

References

[1] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.