

## Uso de funções de similaridade para avaliar a diversidade de conjuntos de testes baseados em modelos.

Eliane Martins, Gustavo L. Machado\*.

### Resumo

Nesse trabalho comparamos os resultados obtidos ao utilizar diferentes funções de similaridade para obter conjuntos de casos de testes baseados em modelos.

### Palavras-chave:

Funções de similaridade, Conjuntos de teste, Análise de agrupamentos.

### Introdução

Geração de casos de teste a partir de modelos de estado costumam produzir um grande volume de dados, ao ponto de poder ser inviável o uso de todos para avaliar o modelo. Mas é intuitivo pensar que dentre toda essa massa de testes existam testes muito parecidos, ou seja, pode ser que utilizar todos os testes, caso possível, não seja uma escolha muito sábia e por isso queremos encontrar formas de otimizar a utilização de casos de teste, possibilitando definir quais casos de testes priorizar.

Dando continuidade ao trabalho no qual escolhemos algumas funções de similaridade para realizar o agrupamento dos casos de teste em conjuntos menores, de acordo com a similaridade entre eles, e também definimos um método para fazer esse agrupamento; vamos agora comparar estas funções.

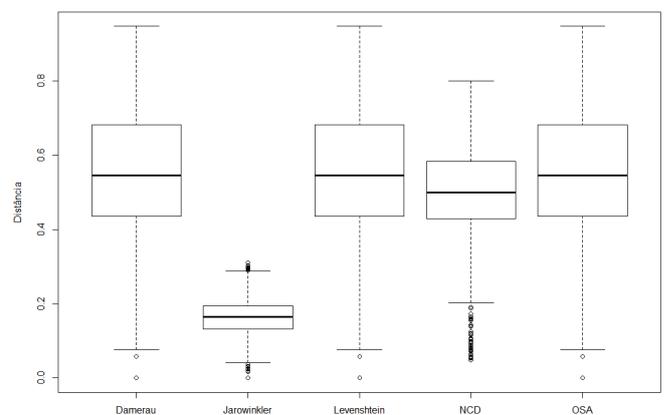
### Resultados e Discussão

Tendo todos os dados para gerar os agrupamentos, criamos boxplots com os conjuntos de dados das distâncias entre os casos de testes geradas pelas diferentes funções de similaridade que utilizamos para termos mais uma base de comparação entre as diferentes distâncias.

Esse tipo de gráfico permite uma análise rápida dos quartis (no primeiro quartil temos 25% dos dados, no segundo 50% e no terceiro 75%), mediana (linha dentro da caixa), diferença entre o primeiro e terceiro quartil (o tamanho da caixa) que representa 50% dos dados e os outliers do conjunto de dados.

| Dados boxplot | LEV  | DL   | OSA  | JW   | NCD  |
|---------------|------|------|------|------|------|
| Min           | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| Q1            | 0.44 | 0.44 | 0.44 | 0.13 | 0.43 |
| Mediana       | 0.55 | 0.55 | 0.55 | 0.16 | 0.50 |
| Média         | 0.55 | 0.55 | 0.55 | 0.16 | 0.50 |
| Q3            | 0.68 | 0.68 | 0.68 | 0.19 | 0.58 |
| Max           | 0.95 | 0.95 | 0.95 | 0.31 | 0.80 |

**Tabela 1.** Dados estatísticos dos conjuntos de distâncias utilizados para montar os boxplots.



**Figura 1.** Boxplot dos valores de distâncias encontrados por cada função de similaridade.

### Conclusões

Observando a Figura 2 é possível ter uma ideia se há ou diferenças entre as distâncias geradas pelas diferentes funções. Caixas sobrepostas implica em uma pequena diferença entre os valores encontrados. Conseguimos notar que as distâncias calculadas pela Jarowinkler (JW) são bem diferentes das demais e também, possui muitos outliers. Assim, pelo fato de ter uma caixa pequena e muitos outliers temos que a JW não é a distância mais adequada para o conjunto de testes que utilizamos.

A função NCD (Normalized Compression Distance) possui muitos outliers, o que não é interessante para nós. Uma primeira análise visual da Figura 2 permite concluirmos que as distâncias Damerau (DAM), Levenshtein (LEV) e OSA são muito parecidas. A análise da Tabela 1 comprova essa análise visual, pois nela conseguimos ver que os dados utilizados para montar o boxplot dessas três funções são idênticos. Isso faz bastante sentido, visto que a DAM e a OSA são variações da LEV.

Esta conclusão do estudo preliminar permitiu que o método a ser utilizado em trabalhos subsequentes fosse estabelecido. O objetivo futuro é aplicar esse método em um conjunto maior de casos de teste.

### Agradecimentos

Ao CNPq pelo apoio financeiro