Data Quality Issues Relating to the NASA Metrics Data Program (MDP) Data Sets

David Gray, David Bowes, Neil Davey, Yi Sun and Bruce Christianson

Science and Technology Research Institute, University of Hertfordshire, UK

January 2010
Introduction & Motivation

- M.Sc. project suggestion - Replicate fault prediction study


- Similar results achieved - **but not identical** . . .

- Motivation to take a closer look at the data.
Data

- Data used originated from NASA.

- Is available online from the NASA Metrics Data Program (MDP) Repository - http://mdp.ivv.nasa.gov/

- Currently has 13 data sets publicly available.

<table>
<thead>
<tr>
<th>v(G)</th>
<th>Arg_Count</th>
<th>No Operands</th>
<th>No Operators</th>
<th>% Comments</th>
<th>LOC_Total</th>
<th>Error_Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
<td>19</td>
<td>44</td>
<td>4</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>51</td>
<td>90</td>
<td>39.22</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>37</td>
<td>74</td>
<td>47.27</td>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>
Issues (1)

- Constant attributes (no information) - 26 attributes KC4
- Repeated / redundant attributes - KC4
- Missing values - 7 / 13 of the data sets

<table>
<thead>
<tr>
<th>No_Lines</th>
<th>Path_Comp</th>
<th>Condition_Count</th>
<th>LOC_Total</th>
<th>Decision_Count</th>
<th>Decision_Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>44</td>
<td>32</td>
<td>16</td>
<td>2.75</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>128</td>
<td>33</td>
<td>54</td>
<td>2.37</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Issues (2)

- Repeated / redundant vectors...

SERIOUS PROBLEM!

Training and testing sets may contain identical vectors!
Issues (3)

Problem domain expertise can help us validate data integrity...

- No. executable lines $< 1$: MC1 4841 vectors (51%)
- No. operators / operands $< 1$: JM1 1332 vectors (12%)
- $v(G) = edges - nodes + 2$ (at the module level...) This does not hold for 145 vectors of MC1 (2%)
Conclusion

There are data quality issues with the NASA MDP data sets…

Of which the repeated / redundant vector issue is most serious.

**Thorough analysis** of your data is essential when data mining.

“it is rather important to explicitly consider the quality – meaning the accuracy – of the data sets that form the basis of our research.” - Liebchen and Shepperd
The End

Thank you for listening... Questions, comments, feedback?
Appendix

First study believed to mention repeated / redundant data in the NASA data sets (and remove it):


First study believed to mention inconsistent data in the NASA data sets: