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

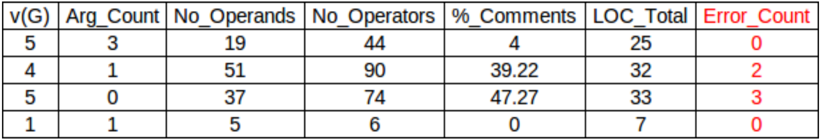
January 2010

Introduction & Motivation

- M.Sc. project suggestion - Replicate fault prediction study
- Menzies et al - *Data Mining Static Code Attributes to Learn Defect Predictors,* IEEE TSE Journal, January 2007
- Similar results achieved - **but not identical. . .**
- Motivation to take a closer look at the data.

Data

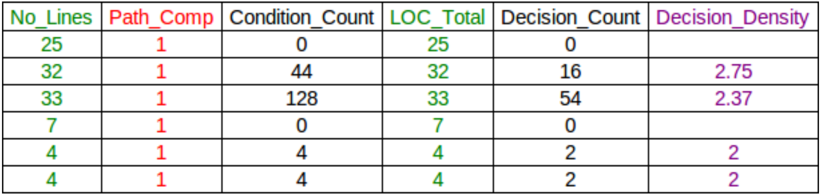
- Data used originated from NASA. . .
- Is available online from the NASA Metrics Data Program (MDP) Repository - http://mdp.ivv.nasa.gov/
- Currently has 13 data sets publicly available. . .



v(G)	Arg_Count	No_Operands	No_Operators	%_Comments	LOC_Total	Error_Count
5	3	19	44	4	25	0
4	1	51	90	39.22	32	2
5	0	37	74	47.27	33	3
1	1	5	6	0	7	0

Issues (1)

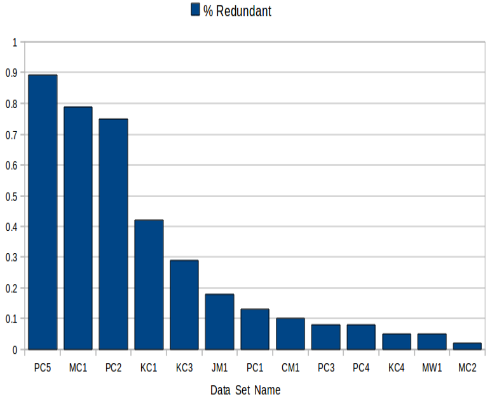
- Constant attributes (no information) - 26 attributes KC4
- Repeated / redundant attributes - KC4
- Missing values - 7 / 13 of the data sets



No_Lines	Path_Comp	Condition_Count	LOC_Total	Decision_Count	Decision_Density
25	1	0	25	0	
32	1	44	32	16	2.75
33	1	128	33	54	2.37
7	1	0	7	0	
4	1	4	4	2	2
4	1	4	4	2	2

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) $= \mathit{edges} - \mathit{nodes} + 2$ (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):

Kaminsky & Boetticher - *Building a Genetically Engineerable Evolvable Program Using Breadth-Based Explicit Knowledge for Predicting Software Defects,* IEEE Annual Meeting of the Fuzzy Information Processing Society, June 2004

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

Seliya, Khoshgoftaar & Zhong - *Analyzing Software Quality with Limited Fault-Proneness Defect Data,* IEEE Int. Symposium on High-Assurance Systems Engineering, October 2005