The scientific basis for prediction research

Martin Shepperd
Dept. of Information Systems & Computing
Brunel University
Uxbridge, UB8 3PH, UK
martin.shepperd@brunel.ac.uk

ABSTRACT

In recent years there has been a huge growth in using statistical and machine learning methods to find useful prediction systems for software engineers. Of particular interest is predicting project effort and duration and defect behaviour. Unfortunately though results are often promising no single technique dominates and there are clearly complex interactions between technique, training methods and the problem domain. Since we lack deep theory our research is of necessity experimental. Minimally, as scientists, we need reproducible studies. We also need comparable studies. I will show through a meta-analysis of many primary studies that we are not presently in that situation and so the scientific basis for our collective research remains in doubt. By way of remedy I will argue that we need to address these issues of reporting protocols and expertise plus ensure blind analysis is routine.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures, process metrics, product metrics*

General Terms

Measurement, Management

Keywords

Software metrics, empirical research, machine learning, defect prediction

1. INTRODUCTION

There has been extensive research over the past 30 or more years into software defect prediction. This has been seen as an important goal since this can assist with both the allocation of testing and verification resources along with decisions concerning the readiness of software for live usage. Most approaches are inductive in the sense that statistical or machine learning methods are used to find predictive models based on attributes derived from static code analysis and/or process metrics such as change data. Useful systematic literature reviews of research progress may be found in [1, 2].

A challenge has been that whilst there has been much research activity and ingenuity in devising new techniques there has been rather less agreement upon the relative effectiveness of these different techniques and clearly no one technique dominates [3]. This inability to replicate results is hindering our ability to make progress and make practical recommendations to practitioners.

Promise '12, September 20–21, 2012, Lund, Sweden

2. META-ANALYSIS

In order to gain a clearer picture of where we as a research community stand I will report on the results of a meta-analysis [**?**] of 601 empirical results derived from *all* relevant primary studies identified by the 2011 Hall et al. systematic review [2]. In the analysis I seek to explore which factors (out of the choice of algorithm, data set, input metrics and research group) have most impact upon results. Surprisingly, the research group is most influential and the choice of algorithm or technique least important. This suggests that researcher bias is confounding our results.

3. SOME REMEDIES

Bias is not a new phenomenon and has been widely reported in other scientific disciplines such as psychology [4] and medicine [5]. The purpose of scientific methods are to reduce bias through the pursuit of transparency and the reduction of subjectivity.

Therefore the question is how should we address this problem. I suggest three courses of action. First, we need better reporting protocols since many machine learning techniques are surprisingly sensitive to different parameter settings and small differences in pre-processing of data sets. Second, we need better sharing of expertise and more joint, prospective primary studies between research groups. Third, we need to make blind analysis the norm, where the nature of the different treatments are hidden from the analyst in order to reduce the subconscious temptation to "cherry pick" results.

Finally, I would note that whilst the focus of my talk is software defect prediction these challenges are relevant to empirical methods in software engineering in general and indeed to the wider scientific community.

Acknowledgements

I would like to thank Tracy Hall and David Bowes for their extensive help in the meta-analysis described in this keynote talk.

4. REFERENCES

[1] C. Catal and B. Diri. A systematic review of software fault prediction studies. *Expert Systems with Applications*, 36(4):7346–7354, 2009.

[2] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell. A systematic review of fault prediction performance in software engineering. *Software Engineering, IEEE Transactions on*, (99):1–1, 2011.

[3] T. Menzies and M. Shepperd. Special issue on repeatable results in software engineering prediction. *Empirical Software Engineering*, pages 1–17, 2012.

[4] C. Mynatt, M. Doherty, and R. Tweney. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *The quarterly journal of experimental psychology*, 29(1):85–95, 1977.

[5] D. Sackett et al. Bias in analytic research. *Journal of chronic diseases*, 32(1-2):51–63, 1979.