Assessing the predictive performance of machine learners in software defect prediction

Martin Shepperd
Brunel University

martin.shepperd@brunel.ac.uk

1

Understanding your fitness function!

Martin Shepperd
Brunel University

martin.shepperd@brunel.ac.uk

2

That ole devil called accuracy (predictive performance)

Martin Shepperd
Brunel University

martin.shepperd@brunel.ac.uk

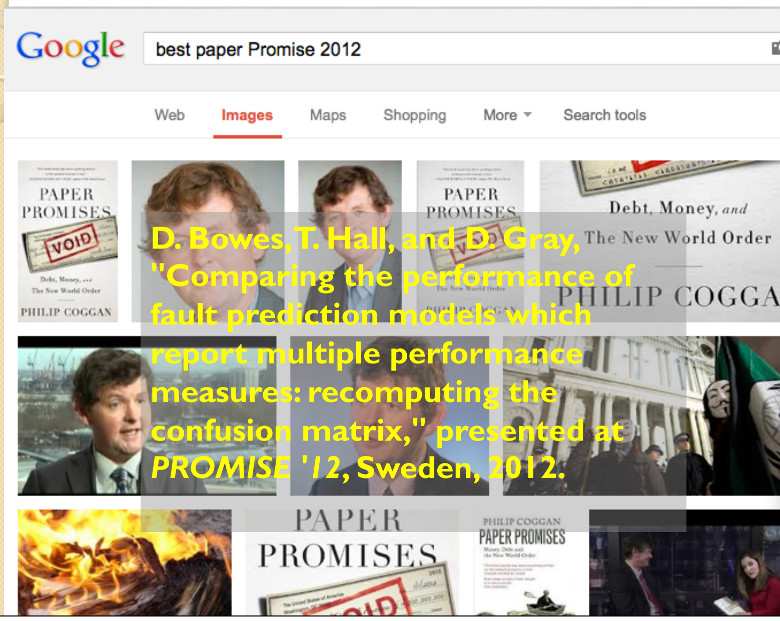
3

Acknowledgements

- Tracy Hall (Brunel)
- David Bowes (Uni of Hertfordshire)

4

Bowes, Hall and Gray (2012)



5

Initial Premises

- lack of deep theory to explain software engineering phenomena
- machine learners widely deployed to solve software engineering problems
- focus on one class – fault prediction
- many hundreds of fault prediction models published [5]

    BUT

- no one approach dominates
- difficulties in comparing results

6

Further Premises

- compare models using prediction performance (statistic)
- view as a fitness function
- statistics measure different attributes / may sometimes be useful to apply multi-objective fitness functions

  BUT!
- need to sort out flawed and misleading statistics

7

Dichotomous classifiers

- Simplest (and typical) case.
- Recent systematic review located 208 studies that satisfy inclusion criteria [5]
- Ignore costs of FP and FN (treat as equal).

- Data sets are usually highly unbalanced i.e., +ve cases < 10%.

8

ML in SE Research Method

```
Invent/find new learner
Find data
REPEAT
  Experimental procedure E yields numbers
  IF numbers from new learner(classifier) >
  previous experiment THEN
    happy
  ELSE
    E' <- permute(E)
UNTIL happy
publish
```


Confusion Matrix

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

- *TP* = true positives (e.g. correctly predicted as defective components)
- *FN* = false negatives (e.g. wrongly predicted as defect-free)
- *TP*, … are instance counts
- *n* = *TP+FP+TN+FN*

Accuracy

$$\frac{TP+TN}{n}$$

- Never use this!
- Trivial classifiers can achieve very high 'performance' based on the modal class, typically the negative case.

11

Precision, Recall and the F-measure

- From IR community
- Widely used
- Biased because they don't correctly handle negative cases.

12

Precision (Specificity)

$$\frac{TP}{TP + FP}$$

- Proportion of predicted positive instances that are correct i.e., True Positive Accuracy
- Undefined if TP+FP is zero (no +ves predicted, possible for *n*-fold CV with low prevalence)

13

Recall (Sensitivity)

$$\frac{TP}{TP + FN}$$

- Proportion of Positive instances correctly predicted.
- Important for many applications e.g. clinical diagnosis, defects, etc.
- Undefined if TP+FN is zero (ie only -ves correctly predicted).

14

F-measure

$$\frac{2 \times R \times P}{R + P}$$

- Harmonic mean of Recall (R) and Precision (P).
- Two measures and their combination focus only on positive examples /predictions.
- Ignores TN hence how well classifier handles negative cases.

$$\begin{array}{cc} \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} & \text{Precision} \\ \text{Recall} & \end{array}$$

15

Different F-measures

- Forman and Scholz (2010)
- Average before or merge?
- Undefined cases for Precision / Recall
- Using highly skewed dataset from UCI obtain F=0.69 or 0.73 depending on method.
- Simulation shows significant bias, especially in the face of low prevalence or poor predictive performance.

16

Matthews Correlation Coefficient

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- Uses entire matrix
- easy to interpret (+1 = perfect predictor, 0=random, -1 = perfectly perverse predictor)
- Related to the chi square distribution

Matthews (1975) and Baldi et al. (2000)

Motivating Example (1)

$$\begin{bmatrix} 10 & 100 \\ 10 & 100 \end{bmatrix}$$

Statistic	Value
n	220
accuracy	0.50
precision	0.09
recall	0.50
F-measure	0.15
MCC	

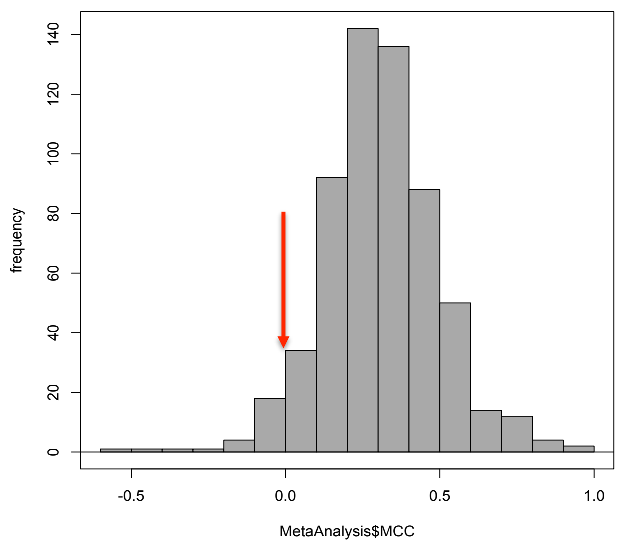
Motivating Example (2)

$$\begin{bmatrix} 10 & 90 \\ 20 & 80 \end{bmatrix}$$

Statistic	Value
n	200
accuracy	0.45
precision	0.10
recall	0.33
F-measure	0.15
MCC	

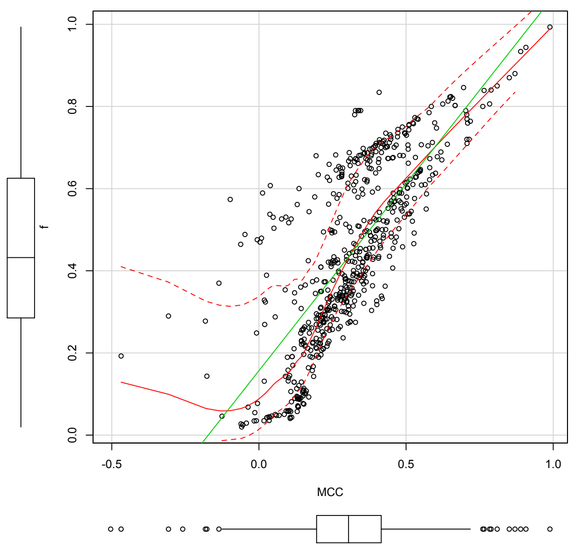
19

Matthews Correlation Coefficient



Martin Shepperd 20

F-measure vs MCC



21

MCC Highlights Perverse Classifiers

- 26/600 (4.3%) of results are negative
- 152 (25%) are < 0.1

- 18 (3%) are > 0.7

Martin Shepperd 22

Hall of Shame!!

a.k.a. accuracy

a.k.a. precision

a.k.a. recall

- The lowest CC valu ually -0.50
- Paper report d:

Table 5: Normalized code vs UML measures

Model	Project	Correctness		Specificity		Sensitivity	
		Code	UML	Code	UML	Code	UML
NRFC	ECS	80%	80%	100%	100%	67%	67%
	CRS	57%	64%	80%	80%	0%	25%
	BNS	33%	67%	50%	75%	0%	50%

- and concluded:

De-spite our encouraging findings, external validity has not been fully proved yet, and further empirical studies are needed, especially with real data from the industry.

Martin Shepperd 23

Hall of Shame (continued)

- A paper in TSE (65 citations) has MCC= -0.47 , -0.31
- Paper reported:

Observed as	Classified as η_1 NFP	η_1 FP	η_2 NFP	η_2 FP	Total
FP	68	20	**75**	13	88
NFP	27	30	23	**34**	57
Total	95	50	98	47	145

- and concluded:

The results show that our approach produces statistically significant estimations and that our overall modeling method performs no worse than existing techniques.

Martin Shepperd 24

Misleading performance statistics

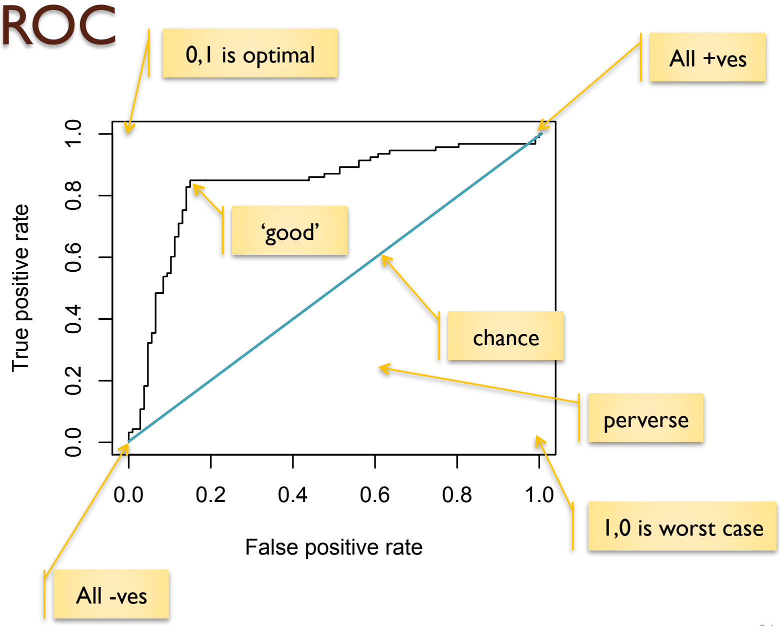
C. Catal, B. Diri, and B. Ozumut. (2007) in their defect prediction study give precision, recall and accuracy (0.682, 0.621, 0.641).

From this Bowes et al. compute an F-measure of 0.6501 [0,1]

But MCC is 0.2845 [-1,+1]

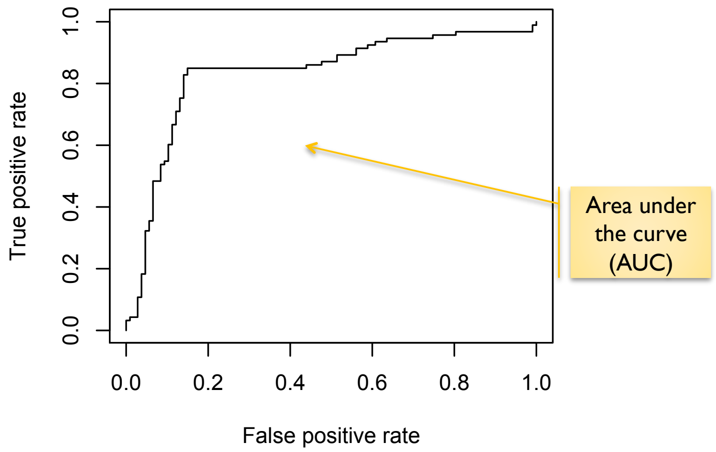
25

ROC



26

Area Under the Curve



27

Issues with AUC

- Reduce tradeoffs between TPR and FPR to a single number
- Straightforward where curve A strictly dominates B -> $AUC_A > AUC_B$
- Otherwise problematic when real world costs unknown

28

Further Issues with AUC

- Cannot be computed when no +ve case in a fold.
- Two different ways to compute with CV (Forman and Scholz, 2010).
  - WEKA v 3.6.1 uses the AUC_{merge} strategy in its Explorer GUI and Evaluation core class for CV, but AUC_{avg} in the Experimenter interface.

29

So where do we go from here?

- Determine what effects we (better the target users) are concerned with? Multiple effects?
- Informs fitness function
- Focus on effect sizes (and large effects)
- Focus on effects relative to random
- Better reporting

30

References

[1] P. Baldi, et al., "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, pp. 412-424, 2000.

[2] D. Bowes, T. Hall, and D. Gray, "Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix," presented at *PROMISE '12*, Lund, Sweden, 2012.

[3] O. Carugo, "Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots," *BMC Bioinformatics*, vol. 8, 2007.

[4] G. Forman and M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, 2010.

[5] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A Systematic Literature Review on Fault Prediction Performance in Software Engineering," IEEE Transactions on Software Engineering, vol. 38, pp. 1276-1304, 2012.

[6] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, pp. 442-451, 1975.

[7] D. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. of Machine Learning Technol.*, vol. 2, pp. 37-63, 2011.

[8] Sing, T., et al., "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, pp. 3940-3941, 2005.

31