A Critique of Software Defect Prediction Models

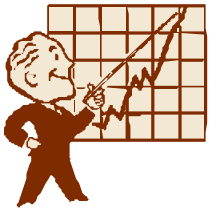
Norman E. Fenton

Dongfeng Zhu

Outline

- What's inside this paper?
- What kind of new technique was developed in this paper?
- Research area of this technique?

My expectation



What's inside this paper?

- Software defect prediction
- Traditional Method
  - Using size and complexity metrics
  - Using Testing Metrics
  - Using process quality data
  - Multivariate approaches
- Bayesian Belief Network (BBN)

Software Defect Prediction

- Defects are commonly defined as deviations from specifications or expectations which might lead to failures in operation.
- Three problem perspectives
  - Predict the number of defects in the system
  - Estimate the reliability of the system in terms of time to failure
  - Understanding the impact of design and testing processes on defect counts and failure densities

Traditional Method

- Complexity and size metrics
- Testing Process
- Design and development process
- Multivariate studies

Size and Complexity

- Lines of Code (LOC)
- McCabe's Cyclomatic Complexity
- Function Point

Lines of Code

Lines of Code (LOC)

D ? 4.86 ? 0.018 L

D ? 0.069 ? 0.00156 L ? 0.00000047 L^2

Lines of Code (LOC)

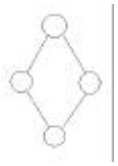
- Strength
  - Can be easily counted
  - Used in many existing software estimation models
- Weaknesses
  - Programming language dependent
  - Smaller programs penalized

McCabe's Cyclomatic Complexity

- Based on a control flow representation of a program. Measure the logical complexity of a module G; Good indicator of module complexity
- Represents the minimum number of linearly independent paths through G
- Cyclomatic complexity is a measure of the amount of control structure or decision logic in a program.
- Study have shown a high correlation between v(G) and the occurrence of errors and it has become a widely accepted indicator of software reliability.

Cyclomatic complexity (cont)

- V = e – n + 2p
  - e is the number of edges
  - n is the number of nodes
  - p is the number of components



Cyclomatic complexity (Cont)

- Strengths
  - Directly comparable across different projects, coding style and languages
  - Used to keep modules within an understandable and workable size
- Weaknesses
  - Formula is incorrect when edges cross
  - Fails to distinguish between different kinds of control flow structures
  - Does not consider the level of nesting of control structure

Function Point

- Function points were introduced in order to replace LOC, provide a more reliable size metric that would capture the added benefit to the user obtained in the developing software, that would be independent of source code and language.
- Function point can be estimated early in the software lifecycle.
- Concept of FP
  - Proposed by Albrecht in 1979
  - UFP =
    number of inputs * w1 + number of outputs * w2
    + number of user inquiries * w3
    + number of files * w4
    + number of external references * w5

Function Point (cont)

weights:

	Simple	Average	Complex
W_1	3	4	6
W_2	3	5	7
W_3	3	4	6
W_4	7	10	15
W_5	5	7	10

Function point counting procedure

- Selection the project to count
- Identify the boundaries of the application
- Select experts
- Calculate the 14 general system characteristic and the value adjustment factor
- Count and weigh
  - EIF (External interface file)
  - ILF (Internal logical file)
  - EI (External inputs)
  - EO (External outputs)
  - EQ (External inquiries)
- Get the unadjusted function point count
- Get the adjusted function point count

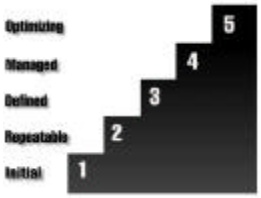
Testing Process

- Prediction based on the collection of data about defects discovered during early inspection and testing phases.
- In the absence of an extensive local database it may be possible to use published bench-marking data to help with this kind of prediction

Process quality Data

- The quality of the development process is the best predictor of product quality.
- The Capability Maturity Model for Software (CMM or SW-CMM) is a model for judging the maturity of the software processes of an organization and for identifying the key practices that are required to increase the maturity of these processes.
- The Software CMM has become a standard for assessing and improving software processes. Through the SW-CMM, the SEI and community have put in place an effective means for modeling, defining, and measuring the maturity of the processes used by software professionals.

Process quality Data (cont)



Need more information, go to :

http://www.sei.cmu.edu/cmm/cmm.html

Multivariate Approaches

- Many of the metrics are colinear
- Design method to choose the principal components
- Develop the relative complexity metric
  - This metric is calculated using the magnitude of variability from each of the factor analysis dimensions as the input weights in a weighted sum

General problems

- Different definition of defects
  - Postrelease defects
  - The total of known defects
  - The set of defects discovered after some arbitrary fixed point in the software life cycle
- The weakness of defect count itself as a measure of software reliability
  - Difficult of determining in advance the seriousness of a defect
  - Wide variations of operational profiles
  - Large numbers of defects may not necessarily lead to improved reliability

Problem with size and complexity method

- Assume a straightforward relationship with defects – defects are a function of size or defects are caused by program complexity
  - The modal ignore the causal effects of programmers and designers
  - Complex program are themselves a consequence of poor design ability or problem difficult
  - Defects may be introduced at the design stage

Problem with Multivariate Approach

- It's hard to see how you might advise a programmer or designer on how to redesign the programs to achieve a better metric

$$control ? a_1 HNK ? a_2 PRC ? a_3 E ? a_4 VG ? a_5 MMC ? a_6 Error ? a_7 HNP ? a_8 LOC$$

Problem with data quality and statistical methodology

- Lack of attention to the assumptions necessary for successful use of a particular statistical technique
- Lack of distinction made between model fitting and model prediction
- Unjustified removal of data points
- Misuse of averaged data

Wake up

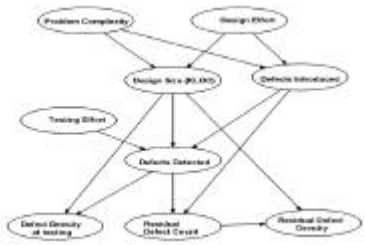
BBN Method

- Bayesian Belief Networks (BBNs) allow us to express complex interrelations within the model at a level of uncertainty commensurate with the problem.
- Bayesian Belief Network, Belief Networks, Causal Probabilistic Networks, Causal Nets, Graphical Probability Networks, Probabilistic Cause-Effect Models, and Probabilistic Influence Diagrams
- A BBN is a graphical network that represents probabilistic relationships among variables. BBNs enable reasoning under uncertainty and combine the advantages of an intuitive visual representation with a sound mathematical basis in Bayesian probability.

BBN Method (cont)

- A BBN is a special type of diagram together with an associated set of probability tables.
- The graph is made up of nodes and arcs where the nodes represent uncertain variables and the arcs the causal/relevance relationships between variables.
- The nodes represent discrete or continuous variables.
- BBN can predict events based on partial or uncertain data. It has the ability to represent and manipulate complex models that might never be implemented using conventional methods

BBN Method (cont)



BBN topology for defect prediction

Bayesian Method

- Bayes set out his theory in 1764
- In making sense of new data, traditional statistical methods ignore the past. Bayesian techniques, in contrast, allow you to start with what you already believe and then see how new information changes your confidence in that belief.
- Putting Bayes theorem to work requires heavy calculus – now aided by computer models – but the fundamental concept is quite straightforward.

Bayesian Method (cont)

$$P(H/D) = \frac{P(H) \cdot P(D/H)}{P(D)}$$

- P(H) prior distribution
- P(D/H) likelihood function
- P(D) evidence distribution
- The theorem says that the probability of the hypothesis, given the data, is equal to the probability of the data, given that the hypothesis is correct, multiplied by the probability of the hypothesis before obtaining the data provided by the averaged probability of the data.

Bayesian Method (cont)

- Advantages
  - Provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence.
  - Bayesian statistics allows researchers to use everything from hunches to hard data to compute the probability that a hypothesis is correct.
  - provides a mechanism for updating state-of-knowledge when the information is accumulated in stages pieces.

Bayesian Method (cont)

- Disadvantage
  - It's difficult to implement before the birth of computers.
  - In theorem, Bayesians say enough evidence will lead people who start with dramatically different priors to essentially the same posterior answer. When the data are scarce, however, the choice of a prior probability can heavily influence the posterior.

BBN Method

- NPT – node probability table
  - The NPTs capture the conditional probabilities of a node given the state of its parent nodes.
- Propagation - propagation algorithms
- Although Bayesian probability been around for a long time it is only in the last few years that efficient algorithms (and tools to implement them) have been developed to enable propagation in networks with a reasonable number of variables
- Hugin tools --- Very powerful BBN tool with excellent graphical interface and fast implementation of Bayesian propagation. Check at www.hugin.com

BBN method (cont)

- Doing propagation is difficult when the number of variable increase or the node dependence increase.
- There is no universal efficient algorithm for doing the computation.
- In the 1980s researchers discovered propagation algorithms that were effective for large classes of BBNs. It is now possible to use BBNs to solve complex problems without doing any of the Bayesian calculations by hand
- The research area in BBN

More resource

- BBN tutorial
  - http://www.dcs.qmul.ac.uk/~norman/BBNs/BBNs.htm
- BBN paper
  - http://www.cs.cmu.edu/afs/cs.cmu.edu/user/ldc/WWW/bayes-net-research/bayes-net-research.html

Thank you!