Empirical Software Engineering

CSE 8340 — Fall 2002

Prof. Jeff Tian, tian@ engr.smu.edu
CSE, SMU, Dallas, TX 75275
(214) 768-2861; Fax: (214) 768-3085
www. engr.smu.edu/~tian/ class/ 8340.02f

Module IIa: Risk Identification

- Risk Identification in ESE
- Techniques and Applications
- Comparison and Recommendations
Risk Identification: Why?

- Observations and empirical evidences:
  - 80:20 rule: non-uniform distribution:
    - 20% of the modules/parts/etc. contribute to
    - 80% of the defects/effort/etc.
  - implication: non-uniform attention
    - risk identification
    - risk management/resolution

- Risk Identification in ESE:
  - 80:20 rule as implicit hypothesis
  - focus: techniques and applications

- Tian SQP paper (survey + comparison) + other technique/application papers
Risk Identification: How?

- Qualitative and subjective techniques:
  - causal analysis
  - Delphi and other subjective methods

- Traditional statistical techniques:
  - correlation analysis
  - regression models:
    - linear, non-linear, logistic, etc.

- Newer techniques:
  - statistical: PCA, DA, TBM
  - AI-based: NN, OSR
  - focus of our study
Risk Identification: Where?

- Characterizing the applications:
  - Goal/target: defect, effort, etc.
    - (80% in the 80:20 rule)
  - Contributor: module, component, etc.
    - (20% in the 80:20 rule)
  - Application domain

- 80% or target:
  - mostly quality or defect
    - (most of our examples also)
  - effort and other external metrics
  - typically directly related to goal
  - resultant improvement
Risk Identification: Where?

● 20% or contributor:
  ▶ 20%: risk identification!
  ▶ understand the link
  ▶ control the contributor:
    – corrections/defect removal/etc.
    – future planning/improvement
    – remedial vs preventive actions

● Application domain:
  ▶ industry: IT, telecom, NASA, etc.
  ▶ process: throughout all phases
  ▶ availability of measurement data
Trad. Tech.: Correlation

- Terminology:
  - r.v.: random variables
  - i.v.: independent (random) variable
    - also called predictor (variable)
  - d.v.: dependent (random) variable
    - also called response (variable)
  - observations and distribution

- Statistical distributions:
  - 1d: normal, exponential, binomial, etc.
  - 2d: independent vs. correlated
  - covariance, correlation (coefficient)
Trad. Tech.: Correlation

- Correlation coefficient:
  - ranges between $-1$ and $1$
  - positive: move in same direction
  - negative: move in opposite direction
  - $0$: not correlated (independent)

- Correlation analysis:
  - use correlation coefficient
  - linear (Pearson) correlation vs. non-parametric (Spearman) correlation
  - based on measurement type/distribution:
    - non-normal distribution
    - ordinal measurement etc.
Trad. Tech.: Correlation

- Correlation analysis: applications
  - understand general relationship
    - e.g., complexity-defect correlation
  - risk identification also
  - cross validation (metrics etc.)

- Correlation analysis: assessment
  - only partially successful
  - low correlation, then what?
  - data skew: 0-defect example
  - uniform treatment of data

- Conclusion: other risk identification techniques are needed.
Trad. Tech.: Regression

- Regression models:
  - as generalized correlation analysis
  - \( n \) i.v. combined to predict 1 d.v.
  - forms of prediction formula
    - \( \Rightarrow \) diff. types of regression models

- Types of regression models:
  - linear: linear function
    \[
    y = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n + \epsilon
    \]
  - log-linear: linear after log-transformation
  - non-linear: non-linear function
  - logistic: represent presence/absence of categorical variables
Trad. Tech.: Regression

- Regression analysis: applications
  - similar to correlation analysis
  - multiple attribute data

- Regression analysis: assessment
  - only partially successful
  - similar to correlation analysis
  - often marginally better (R-sqr vs c.c.)
  - same kind of problems
  - data transformation problem
  - synthesized metrics $\sim$ regression model?

- Conclusion: other risk identification techniques are often needed.
New Techniques

- New statistical techniques:
  - PCA: principal component analysis
  - DA: discriminant analysis
  - TBM: tree-based modeling

- AI-based new techniques:
  - NN: artificial neural networks
  - OSR: optimal set reduction
  - others: pattern matching, abductive reasoning, etc.

- Focus of our study: rest 4 papers.
New Techniques: PCA & DA

- PCA: principal component analysis
  - idea of linear transformation
  - PCA to reduce dimensionality
  - effective in combination with other techniques: DA etc.

- DA: discriminant analysis
  - discriminant function
  - combine with other techniques

- Not really new techniques, but rather new applications in SE.

- 1/1996 Paper by Khoshgoftaar et al., IEEE Software
New Techniques: TBM

- TBM: tree-based modeling
  - similar to decision trees
  - but data-based (derived from data)
  - preserves tree advantages:
    - easy to understand/interpret
    - both numerical and categorical data
    - partition \(\Rightarrow\) non-uniform treatment

- TBM applications:
  - main: defect analysis
  - past: psychology, SE-Amadeus, etc.
  - reliability: TBRMs
  - expertise/experience from SMU team

- 2001 Paper by Tian et al., JSS 57(3)
New Techniques: NN

- NN: artificial neural networks
  - NN or ANN inspired by biological computation
  - neuron: basic computational unit
    - different functions
  - connection: neural network
  - input/output/hidden layers

- NN applications:
  - AI and AI problem solving
  - main: defect/risk identification

- 1995 Paper by Khoshgoftaar et al., ASE
New Techniques: OSR

- OSR: optimal set reduction
  - pattern matching idea
  - clusters and cluster analysis
  - similar to TBM but different in:
    - pattern extraction vs. partition
  - original paper by Briand et al., 1992

- OSR applications: 1993 Paper by Briand/Basili/Hetmanski, TSE 19(11)

- Other AI/pattern matching techniques:
  - rich literature
  - few applications
Risk Identification: Comparison

- Apply the GQM paradigm.

- Goals of comparison:
  - risk identification goal?
  - how effective in achieving this goal
  - relate to general modeling/analysis goals:
    - assessment/prediction/control
    - focus on control or improvement

- General questions:
  - assessment: accurate, over time
  - prediction: accurate and early
  - control: many aspects
    - useful information and guidance
    - easy to use for improvement actions
    - tracking improvement over time
Risk Identification: Comparison

- Specific questions for comparison:
  - accuracy
  - simplicity
  - early availability and stability
  - ease of result interpretation
  - constructive information and guidance for (quality) improvement
  - availability of tool support

- Comparison (metric):
  - qualitative based on above questions
  - quantitative in the future?

- Recommendation based on comparison and application environment.
Comparison: Accuracy

- Accuracy in assessment:
  - model fits data well
    - use various goodness-of-fit measures
  - avoid over-fitting
  - cross validation by review etc.

- Accuracy in prediction:
  - over-fitting ⇒ bad predictions
  - prediction: training and testing sets
    - within project: jackknife
    - across projects: extrapolate
  - prediction errors: type I & II

- General comparison here.
  (More quantitative comparison later.)
Comparison: Usability

- Can’t explain in a few words
  ⇒ difficulties with reception/deployment

- Simplicity & result interpretation?
  - technique easy to use/understand
  - what does it (the result) mean?
  - training effort involved
  - causal and other connections

- Tool and other support:
  - availability of easy-to-use tools
  - other support: process/personnel/etc.
  - direct impact on deployment
Comparison: Usefulness

- Early availability and stability
  - to be useful must be available early
  - focus on control/improvement
  - apply remedial/preventive actions early
  - track progress: stability

- constructive information and guidance
  - what: assessment/prediction
  - how to improve?
    - constructive information
    - guidance on what to do
  - example of TBRMs
Comparison Summary & Perspectives

• Summary of comparison:
  ▶ Figure 11.
  ▶ examples, and more examples to follow
  ▶ TBM good balance

• Recommendation:
  ▶ qualitative based on above questions
  ▶ quantitative in the future?

• Integration into SE Process