Brunel University London

~~Bayesian Modeling for~~ Software Quality
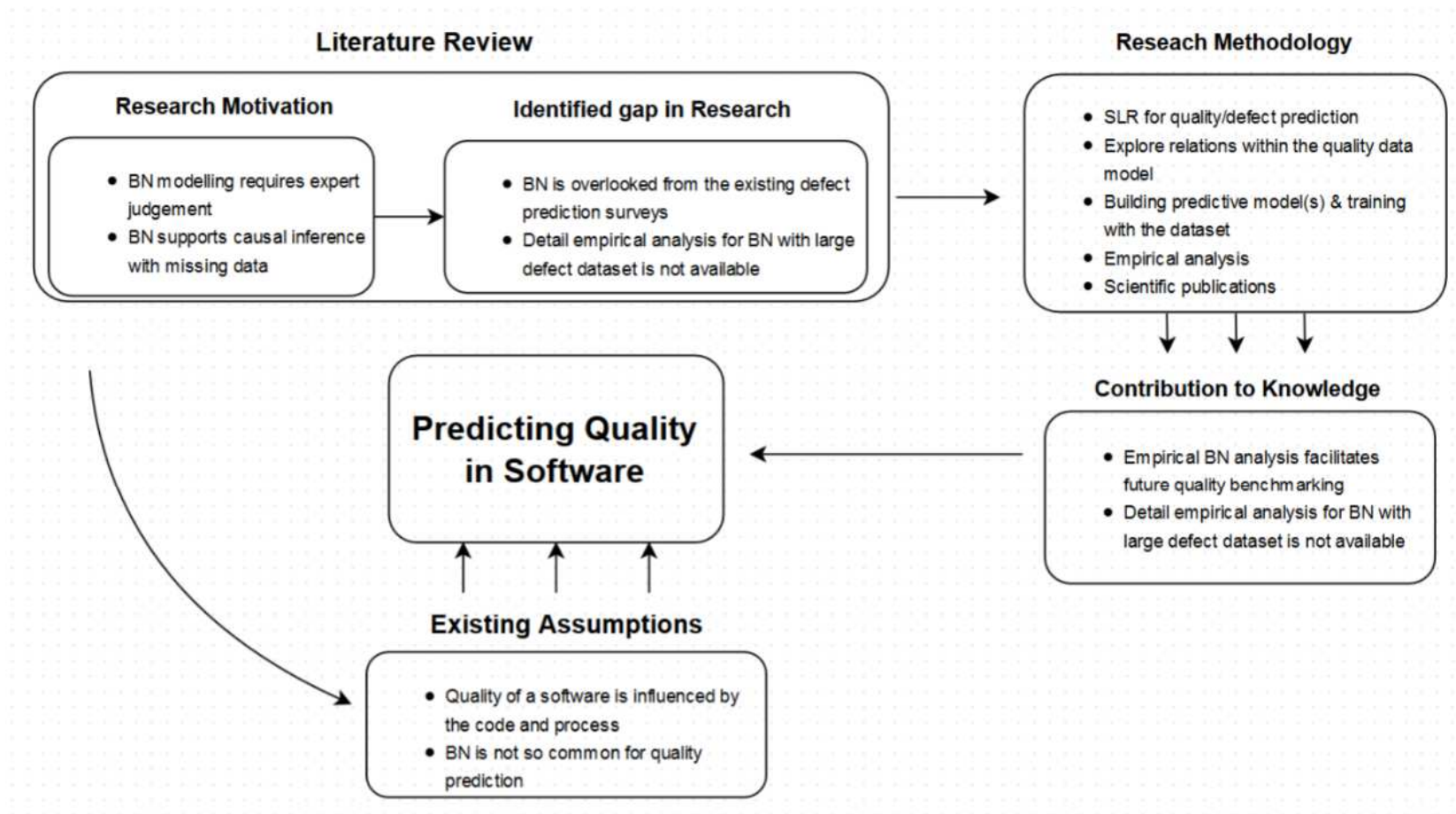
Data Mining to Predict

By

Mohammad Sharif Hanif

Department of Computer Science
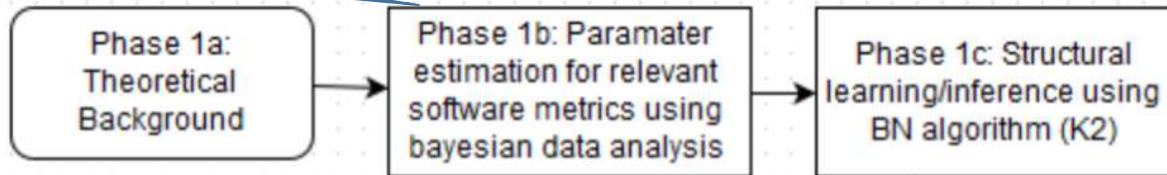
Brunel University London

06 July 2015

Literature Review

Reseach Methodology

Research Motivation

- BN modelling requires expert judgement
- BN supports causal inference with missing data

Identified gap in Research

- BN is overlooked from the existing defect prediction surveys
- Detail empirical analysis for BN with large defect dataset is not available

- SLR for quality/defect prediction
- Explore relations within the quality data model
- Building predictive model(s) & training with the dataset
- Empirical analysis
- Scientific publications

Contribution to Knowledge

- Empirical BN analysis facilitates future quality benchmarking
- Detail empirical analysis for BN with large defect dataset is not available

Predicting Quality in Software

Existing Assumptions

- Quality of a software is influenced by the code and process
- BN is not so common for quality prediction

Conceptual Framework for the research

Research Phases

Classification

Count
10686
437

| Phase 1a: Theoretical Background | → | Phase 1b: Paramater estimation for relevant software metrics using bayesian data analysis | → | Phase 1c: Structural learning/inference using BN algorithm (K2) |

K2 search & Association rule mining

Count
15954
13539

| Phase 2b: Sub-class discovery of faults | ← | Phase 2a: Component Prediction |

Clustering
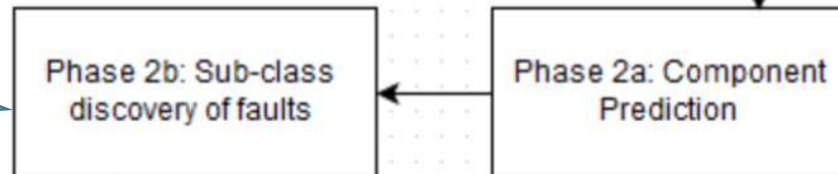
Text mining

Relation: textmining

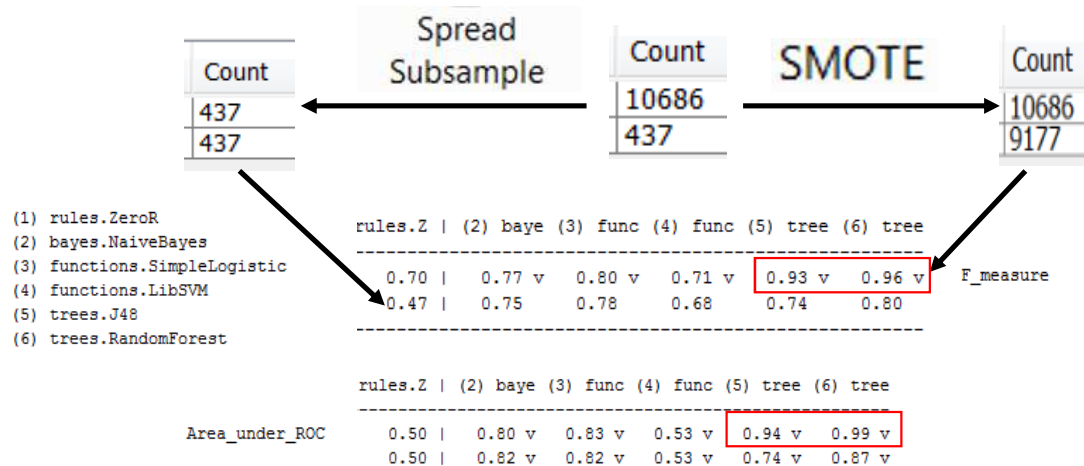| No.	text (String)	category (Nominal)
1	c client: use Wdeclarationafterstatement Visual studio still does...	0
2	ACLs are not respected for node deletion Any session can delete...	1
3	Quota is not correctly rehydrated on snapshot reload traverseNo...	1
4	setup the new website on zookeeperapacheorg This uses the ne...	0
5	Replace timers with semaphores in FLENewEpochTest This patc...	0
6	fn:error missing The fn:error function has not been built	0
7	LogFormatter should support printing transaction data as a strin...	1

| Phase 3a: Improving the predictive performance with Active Learning | → | Phase 3b: Thesis writing and Graduation |

Unsupervised learning
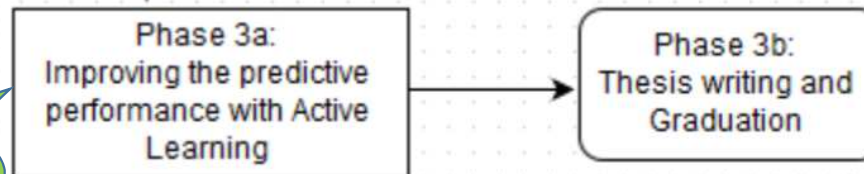
| Color Code | TODO | WIP | DONE |

Data Mining for Rare Class Analysis

Spread Subsample	Count		SMOTE	Count
	10686			10686
	437			9177

Count
437
437

(1) rules.ZeroR
(2) bayes.NaiveBayes
(3) functions.SimpleLogistic
(4) functions.LibSVM
(5) trees.J48
(6) trees.RandomForest

rules.Z	(2) baye	(3) func	(4) func	(5) tree	(6) tree	
0.70	0.77 v	0.80 v	0.71 v	0.93 v	0.96 v	F_measure
0.47	0.75	0.78	0.68	0.74	0.80	

	rules.Z	(2) baye	(3) func	(4) func	(5) tree	(6) tree
Area_under_ROC	0.50	0.80 v	0.83 v	0.53 v	0.94 v	0.99 v
	0.50	0.82 v	0.82 v	0.53 v	0.74 v	0.87 v

=== Summary ===

Correctly Classified Instances	19424	97.7899 %
Incorrectly Classified Instances	439	2.2101 %
Kappa statistic	0.9555	
Mean absolute error	0.0713	
Root mean squared error	0.1473	
Relative absolute error	14.3497 %	
Root relative squared error	29.5451 %	
Total Number of Instances	19863	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.98	0.024	0.979	0.98	0.979	0.996	0
	0.976	0.02	0.976	0.976	0.976	0.996	1
Weighted Avg.	0.978	0.022	0.97	0.978	0.978	0.996	

=== Confusion Matrix ===

```
    a      b   <-- classified as
 10467   219 |    a = 0
   220  8957 |    b = 1
```

Keywords

Cost Sensitive Learning

One Class Learning with Positive Class

Weighting Methods

Selected attributes: 1,8,9,10,15,18,19 : 7
```
     id
     HBUG
     CREF
     XMET
     MOD
     EXCT
     EXCR
```
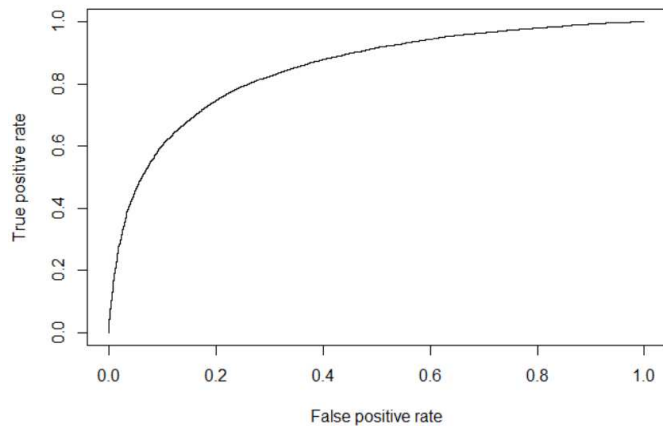
Ranked attributes:
```
  0.6032    9 CREF
  0.5893    3 NOCL
  0.5535    4 NOS
  0.5509   24 VREF
  0.5506    6 HVOC
```

=== Confusion Matrix ===

```
    a      b   <-- classified as
 10345   341 |    a = 0
   232  8945 |    b = 1
```

=== Confusion Matrix ===

```
    a      b   <-- classified as
 10518   168 |    a = 0
   339  8838 |    b = 1
```

Phase 2a. Text Mining

Current relation
    Relation: textmining-weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-prune-rate-1.0-N0-stemmerweka.core.ste...
    Instances: 29493   Attributes: 1229

Feature selection with the most predictive variables (here words)

```
Selected attributes: 20,25,81,122,147,150,325,371,409,433,434,541,655,874,917,978,997,1062,1227 : 19
                     Add
                     Allow
                     Implement
                     Provide
                     Support
                     The
                     correctly
                     does
                     error
                     fail
                     fails
                     is
                     not
                     support
                     to
                     when
                     would
                     broken
                     wrong
```

```
> print(attributes(performance(eval,'auc'))$y.values[[1]])
[1] 0.8478918


> #distribution of bug and not-bug
> table(report.raw$category)

    0     1
15954 13539
```

sparse matrix with most frequent words that appear at 10 times in the corpus.

Next task: ?

Next meeting: ?