Assignment 2

Temesgen Mehari, Oliver Philipps, Stefan Vikoler

Data Engineering

09. December 2015

Task I: Business understanding

- Business objectives
  - What is the problem?
    - Predict defects in software.
    - Reduce support events.
    - Higher quality in software.
  - Who has the problem?
    - Product managers and developers.
  - Problem measurement
    - Amount of defects in the software (support events).
- Business success criteria
  - Better quality in software.
  - Increase customer satisfaction.
  - Better reputation.
  - Increase competition on the market and get new customers.
  - Earn more money.
  - Reduce costs for support.

Task I: Business understanding

- Determine data mining goals
  - Find critical software metrics.
  - Estimate how likely an existing product is faultless.
  - Define quality of software.
- Data mining success criteria
  - Estimated defect and the real defect
  - Correctly Classified Instances $> 80\%$

Task II: Data understanding

- 10885 instances with 22 fields
  - 21 numeric software metrics
  - 1 boolean 'defect' {true,false}
- 2106 false (19,35%), 8779 true (80,65%)
- A lot of high correlation
  - i.e. loc has high correlation with v(g), iv(g), n, v, e, b, t, lOCode, lOBlank, uniq_Opnd, total_Op, total_Opnd, branchCount

Task III: Data preparation

- Numerics
    - false = 0 and true = 1
- Balancing
    - 1400 false, 1400 true instances for training
    - 700 false, 700 true instances for testing
- Training and testing data set
    - $\frac{2}{3}$ training
    - $\frac{1}{3}$ testing
- Remove NULL or unusable values

Task IV: Modeling

- WEKA Logistic function

$$\frac{1}{1 + \sum_{j=1}^{k-1} e^{X_i \cdot B_j}}$$

- Logistic regression with whole dataset: Correctly Classified Instances of 81.8968%

Task V: Evaluation

Correctly Classified Instances	806	57.5714%
Incorrectly Classified Instances	594	42.4286%
Mean absolute error	0.4709	
Root mean squared error	0.4952	
Relative absolute error	94.1893%	
Root relative squared error	99.0445%	
Total Number of Instances	1400	

Task V: Evaluation

a	b	← classified as
520	180	a = false
414	286	b = true

Task VI: Plan deployment

- Development process
- Measure
  - at the end of development
  - amount of defects
  - amount of support events
  - customer feedback
- Propagate the knowledge
  - Kick off meeting
  - meetings on a regular base (i.e. daily scrum)
- Pitfalls
  - Costs of time for the developers.