Categorical missing data imputation for software cost estimation by multinomial logistic regression

Panagiotis Sentas and Lefteris Angelis
The Journal of Systems and Software, 2006

2008. 4. 16
Yeong-Seok Seo
Contents

- Introduction
- Experimental design: Overview
  - Step 1. Data preprocessing
  - Step 2. Simulation of missing values
  - Step 3. Missing data techniques
  - Step 4. Results and analysis
- Related work
- Conclusion
- Discussion
Attempts to estimate the cost usually begins by building estimation model

- Apply estimation method to historical data sets with completed software project data

A major problem to build a estimation model is the missing data in software project data

- Mislead the models’ accuracy by the lack of data values in several important project variables
Imputation methods for missing data are mainly applied on numerical data
- Contain mostly categorical variables with many missing values in software project data sets

Investigate the possibility of using multinomial logistic regression as imputation method for categorical variable
- Compared with four other imputation methods based on the estimation accuracy of a regression model after applying each method
Experimental design: Overview

Step 1. Data preprocessing
- ISBSG – R7
- 1,238 projects
- 166 projects

Step 2. Simulation of missing values
- 1) MCAR
- 2) MAR
- 3) NIM
- 1) 10%
- 2) 20%
- 3) 30%

Step 3. Missing data techniques
- (1) Mean imputation
- (2) Listwise deletion
- (3) Expectation maximization
- (4) Regression imputation
- (5) Multinomial logistic regression

Step 4. Results and analysis

© KAIST SE LAB 2008
Step 1. Data preprocessing (1/2)

- ISBSG (Release 7) – 1,238 projects
  - Selected projects
    - 166 projects remained for this study (Training: 150, Testing: 16)
      - Select high data quality ratings
      - Remove projects with missing values
  - Selected variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full name</th>
<th>Levels-definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>size</td>
<td>Project size</td>
<td>Numeric</td>
</tr>
<tr>
<td>develop</td>
<td>Development type</td>
<td>Enhancement, New Development, Re-development</td>
</tr>
<tr>
<td>platfr</td>
<td>Development platform</td>
<td>MF, MR, PC</td>
</tr>
<tr>
<td>lang</td>
<td>Language type</td>
<td>3GL, 4GL, ApG</td>
</tr>
<tr>
<td>primar</td>
<td>Primary programming language</td>
<td>ACCESS, C, C++, CLIPPER, COBOL, COBOL II, VB, …</td>
</tr>
<tr>
<td>implem(year)</td>
<td>Implementation date</td>
<td>Numeric</td>
</tr>
<tr>
<td>orgtype</td>
<td>Organization type</td>
<td>Banking, Construction, Gas, Defense, Energy, …</td>
</tr>
<tr>
<td>bartype</td>
<td>Business area type</td>
<td>Accounting, Legal, Insurance, …</td>
</tr>
<tr>
<td>apltype</td>
<td>Application type</td>
<td>Network management, Decision support system, …</td>
</tr>
<tr>
<td>pacost</td>
<td>Package customization</td>
<td>Don’t know, Yes, No</td>
</tr>
</tbody>
</table>
Step 1. Data preprocessing (2/2)

- ISBSG (Release 7) – 1,238 projects (cont’d)

  Merging the original categories into homogeneous groups because of a large number of categories

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full name</th>
<th>Description and levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>develop</td>
<td>Development type</td>
<td>1=Enhancement, 2=New Development, 3=Re-development</td>
</tr>
<tr>
<td>platfr</td>
<td>Development platform</td>
<td>1=MF, 2=MR, 3=PC</td>
</tr>
<tr>
<td>lang</td>
<td>Language type</td>
<td>1=3GL, 2=4GL, 3=ApG</td>
</tr>
<tr>
<td>primar_4</td>
<td>Primary programming language</td>
<td>1 = {access}, 2 = {easytrieve, natural, oracle, pl/i, power builder, visual basic}, 3 = {cobol, ideal, other apg, sql, telon}, 4 = {C,C++, clipper, cobol II, other 4gl}</td>
</tr>
<tr>
<td>orgtype_4</td>
<td>Organization type</td>
<td>1 = {communication, computers, consultancy, energy, financial,…}, 2 = {banking, community services, construction, defence, electronics,…}, 3 = {electricity, gas, water}, 4 = {aerospace/automotive, consumer goods, distribution, government,…}</td>
</tr>
<tr>
<td>bartype_4</td>
<td>Business area type</td>
<td>1 = {activity tracking, research and development, telecommunications,…}, 2 = {engineering}, 3 = {accounting, banking, financial (excluding banking),…}, 4 = {architectural, blood bank, insurance, inventory, legal, manufacturing,…}</td>
</tr>
<tr>
<td>apltype_4</td>
<td>Application type</td>
<td>1 = {decision support system, network management, transportation,…}, 2 = {management information system, process control}, 3 = {inventory control, system conversion, transaction/production system}, 4 = {office information system, technical information system}</td>
</tr>
</tbody>
</table>
### Three missing data mechanisms

<table>
<thead>
<tr>
<th>Missing completely at random (MCAR)</th>
<th>• The missing values are unrelated to the values of any other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating missing values completely at random for each variable</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Missing at random (MAR)</th>
<th>• The probability of having missing values depend on the values of some other variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending order of size</td>
<td><img src="image" alt="Diagram" /></td>
</tr>
<tr>
<td>Sort according to the same variable</td>
<td><img src="image" alt="Diagram" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-ignorable missingness (NIM)</th>
<th>• The probability of having missing values depend on the variable itself</th>
</tr>
</thead>
</table>

(p: Total number of missing data)
Step 2. Simulation of missing values (2/2)

- Three percentages of incomplete data
  - 10%, 20%, and 30%, which are quite realistic for any data set

1. MCAR
2. MAR
3. NIM

- 1) 10%
- 2) 20%
- 3) 30%

1> platfr
2> primar_4
3> bartype_4
4> apltype_4
Step 3. Missing data techniques (1/3)

- **Five techniques used in this paper**
  - **Listwise deletion (LD)**
    - Delete the data with missing values
    - Simplicity, but cause loss of precision and bias
  - **Mean imputation (MI)**
    - Replace the missing values with the mean of the observed values in that variable
    - Generally perform well, when the data are normal distribution
  - **Regression imputation (RI)**
    - Estimate the missing values through the application of multiple regression
Step 3. Missing data techniques (2/3)

- **Five techniques used in this paper (cont'd)**
  - **Expectation maximization (EM)**
    - Replace the missing value through the distribution for the missing data based on the known values
  - **Multinomial logistics regression (MLR)**
    - Finds a probability for each category value, then imputes a value using those probabilities
    - Need $q-1$ functions if the possible categories are $q$

$$
\log \left( \frac{\text{prob}(\text{category}_j)}{\text{prob}(\text{category}_q)} \right) = b_0^{(j)} + \sum_{i=1}^{k} b_i^{(j)} x_i \quad (j = 1, 2, ..., q-1)
$$
Step 3. Missing data techniques (3/3)

- Five techniques used in this paper (cont’d)

1. MCAR
   - 1) 10%
   - 2) 20%
   - 3) 30%

2. MAR
   - 1> platfr
   - 2> primar_4
   - 3> barytype_4
   - 4> apltype_4

3. NIM

(1) Mean imputation
(2) Listwise deletion
(3) Expectation maximization
(4) Regression imputation
(5) Multinomial logistic regression
Step 4. Results and analysis (1/9)

- **Comparison criterion**
  - Assess the estimation accuracy of an effort estimation model
    - The measure of accuracy used is the standard deviation (SD) of the residual error
    - Small values of SD indicate high estimation accuracy

\[
SD = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-1}}
\]

- \( y_i \): Actual Effort
- \( \hat{y}_i \): Predicted Effort
- \( n \): Size of the test set
Step 4. Results and analysis (2/9)

- Overall performance of the MDTs
  - Descriptive statistics
    
    | Method | Mean      | Median   | SD           |
    |--------|-----------|----------|--------------|
    | SD     | 2985.328  | 2961.524 | 269.6245892  |
    | LD     | 2983.820  | 2986.307 | 353.5260966  |
    | EM     | 2942.735  | 2964.400 | 214.8482471  |
    | RI     | 2801.121  | 2800.546 | 255.0356807  |
    | MLR    | 2729.260  | 2746.551 | 178.4105630  |

  - Smaller mean, median, SD

- Box-plots
Step 4. Results and analysis (3/9)

- Mean performance for three different percentages
  - 10%, 20%, 30%

![Graphs showing performance at 10% and 30%](image-url)
Step 4. Results and analysis (4/9)

- Mean performance for three different mechanisms
  - MCAR, MAR, NIM

![Graphs showing estimated marginal means for different methods under MCAR and NIM mechanisms.](image)
Step 4. Results and analysis (5/9)

- **Mean performance for four categorical variables**
  - Each line corresponds to a mechanism

![Graphs showing mean performance for four categorical variables](image)
Mean performance for four categorical variables
- Each line corresponds to a mechanism
Step 4. Results and analysis (7/9)

- Mean performance for four categorical variables
  - Each line corresponds to a percentage level

(a) At Variable = bartype_4
(b) At Variable = primar_4

© KAIST SE LAB 2008
Step 4. Results and analysis (8/9)

- Mean performance for four categorical variables
  - Each line corresponds to a percentage level

(c) At Variable = aptype_4

(d) At Variable = platfr

Estimated Marginal Means

(c) Method

(d) Method
## Ranking of the methods performance

### Number of times that each MDT gave the best accuracy

<table>
<thead>
<tr>
<th></th>
<th>MI</th>
<th>LD</th>
<th>EM</th>
<th>RI</th>
<th>MLR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAR</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIM</td>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>20% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAR</td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIM</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>30% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAR</td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIM</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Ranking score of each MDT

<table>
<thead>
<tr>
<th></th>
<th>MI</th>
<th>LD</th>
<th>EM</th>
<th>RI</th>
<th>MLR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>MAR</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>NIM</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td><strong>20% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>MAR</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>NIM</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td><strong>30% of missing data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCAR</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>MAR</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>NIM</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>14</td>
</tr>
</tbody>
</table>

Both tables show the superiority of MLR
# Related work

<table>
<thead>
<tr>
<th>Authors</th>
<th>Missing data techniques</th>
<th>Missing data mechanisms</th>
<th>Data set</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strike et al.</td>
<td>LD, MI, 8 different types of hot-deck imputation</td>
<td>MCAR, MAR, NIM</td>
<td>206 software projects from 26 different companies</td>
<td>Prediction accuracy of software effort estimation model</td>
</tr>
<tr>
<td>Myrtveit et al.</td>
<td>LD, MI, SRPI, FIML</td>
<td>MCAR, MAR</td>
<td>176 ERP projects</td>
<td></td>
</tr>
<tr>
<td>Cartwright et al.</td>
<td>SMI, K-NN</td>
<td>Real missing data</td>
<td>17 Bank data, 21 multi-national data</td>
<td></td>
</tr>
<tr>
<td>Song et al.</td>
<td>CMI, K-NN</td>
<td>MCAR, MAR</td>
<td>50 and 100 cases from ISBSG release 7</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

**Contribution**

- Investigate the use of MLR for estimating the missing values of categorical variables
  - MLR performs better or at least equally well
  - LD is the most popular one, however it causes loss of precision and bias

**Future work**

- MLR compares with other methods
  - Real missing categorical data, especially in the full ISBSG
  - Different effort estimation models
  - Other application domains in software engineering
Discussion

- One issue is the generation of the missing values
  - Missing values appear across all of the variables in real data sets

- MDL is a purely categorical imputation method
  - Categorical values have a little effect on the effort estimation model

* We had a two data set !!!
- The results of the missing data imputation can be compared