Combining techniques to optimize effort predictions in software project management


Stephen G. MacDonell and Martin J. Shepperd

2008. 8. 14
Sunkyung Lee
Contents

- Introduction
- Background
- Prediction techniques
  - Expert judgment
  - Least squares linear regression
  - Case-based reasoning
- Overall Approach
- Case study
- Conclusion
Current issue
- Many prediction techniques are suggested
- None of prediction technique has proved consistently accurate
- Some authors have suggested the use of multiple prediction techniques

Motivation
- There is little empirical work using multiple techniques
- Empirical investigation is needed to find the efficiency of multiple techniques
Research goal

- Prove the potential benefit of multiple techniques
- Find the rule which prediction technique to use in which conditions
Background(1/4)

- Mean Magnitude of Relative Error (MMRE)
  - Measures of the estimation accuracy
  - Lower MMRE, better accuracy

\[
MRE = \frac{|V_a - V_p|}{|V_a|} \quad \text{absolute error}
\]

\[
MMRE = \frac{1}{n} \sum_{j=1}^{n} MRE_j
\]

\(V_a\): actual value
\(V_p\): predicted value
\(n\): total number of data points

<table>
<thead>
<tr>
<th>Actual value</th>
<th>Predicted value</th>
<th>Absolute error</th>
<th>Relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
<td>0.2857</td>
</tr>
<tr>
<td>457</td>
<td>345</td>
<td>112</td>
<td>0.2451</td>
</tr>
</tbody>
</table>
Covariance

A measure of how much two variables change together

\[ \text{Cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} \]  

(E(X), \mu_x: mean)

Higher(positive) 0 Higher(Negative)
Correlation

- A measure of strength of a linear relationship between two variables
- Upper bound: 1 / Lower bound: -1
- Revision of covariance about unit
- Inter-item correlation
  - Correlation between items in a measure

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>kg</td>
<td>cm</td>
<td>Different result</td>
</tr>
<tr>
<td>kg</td>
<td>Inch</td>
<td></td>
</tr>
</tbody>
</table>

$1 \text{ cm} = 0.393700787 \text{ inch}$

$\Rightarrow$ Used to find the relationship between dependent variables and independent variables
Rule induction (RI) algorithm C5.0*

- Find small (or simple) decision tree
  - Learn the classification rule of class

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play / Don’t Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>85</td>
<td>85</td>
<td>FALSE</td>
<td>Don’t Play</td>
</tr>
<tr>
<td>sunny</td>
<td>80</td>
<td>90</td>
<td>TRUE</td>
<td>Don’t Play</td>
</tr>
<tr>
<td>sunny</td>
<td>72</td>
<td>95</td>
<td>FALSE</td>
<td>Don’t Play</td>
</tr>
<tr>
<td>sunny</td>
<td>69</td>
<td>70</td>
<td>FALSE</td>
<td>Play</td>
</tr>
<tr>
<td>sunny</td>
<td>75</td>
<td>70</td>
<td>TRUE</td>
<td>Play</td>
</tr>
<tr>
<td>overcast</td>
<td>83</td>
<td>78</td>
<td>FALSE</td>
<td>Play</td>
</tr>
<tr>
<td>overcast</td>
<td>64</td>
<td>65</td>
<td>TRUE</td>
<td>Play</td>
</tr>
<tr>
<td>overcast</td>
<td>72</td>
<td>90</td>
<td>TRUE</td>
<td>Play</td>
</tr>
<tr>
<td>overcast</td>
<td>81</td>
<td>75</td>
<td>FALSE</td>
<td>Play</td>
</tr>
<tr>
<td>rain</td>
<td>65</td>
<td>70</td>
<td>TRUE</td>
<td>Don’t Play</td>
</tr>
<tr>
<td>rain</td>
<td>71</td>
<td>80</td>
<td>TRUE</td>
<td>Don’t Play</td>
</tr>
<tr>
<td>rain</td>
<td>68</td>
<td>80</td>
<td>FALSE</td>
<td>Play</td>
</tr>
<tr>
<td>rain</td>
<td>75</td>
<td>80</td>
<td>FALSE</td>
<td>Play</td>
</tr>
<tr>
<td>rain</td>
<td>70</td>
<td>96</td>
<td>FALSE</td>
<td>Play</td>
</tr>
</tbody>
</table>

* C5.0 : commercial version of C4.5
  - Speed : C5.0 is faster than C4.5
  - Memory : C5.0 is more memory efficient
  - Smaller Decision Tree
    : C5.0 has smaller Decision Trees
Prediction techniques (1/3)

- Expert judgment
  - Rely heavily on the experience of expert’s knowledge in
    - Similar development environments
    - Historically maintained DB on completed project
    - Accuracy of these past project
  - Commonly used, but there’s evidence of bias*

* Factors systematically associated with errors in subjective estimates of software development effort (Gray et al., 1999)
Least squares linear regression (LSR)

- Find a straight-line relationship between parameters and dependent variables
  \( \hat{Y} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n \) (\( \hat{Y} \): predicted value of \( Y \))

- Vulnerable to extreme outlier and collinearity
  - Collinearity indicates that a set of points on a single line
    - Independent variables have strong linear relationship each other
  - It’s difficult to use LSR when data set have collinearity

  ex) \( \hat{Y} = \beta_0 + \beta_1X_1 + \beta_2X_2 \)
  \( X_2 = 2.4X_1 \)
Prediction techniques (3/3)

- **Case-based reasoning (CBR)**
  - Find the similar cases to the target project
  - The most similar project’s effort value is used to predict effort
  - More robust to problems than LSR
Overall approach

Medical record Database system

Comparison of accuracy
(Absolute residual, MMRE)

Find the potential benefit of using multiple techniques
(Covariance)

Expert judgment
LSR
CBR

Expert judgment
LSR
CBR

Single techniques
LSR

Multiple techniques
Average from three single techniques
Rule induction

Attempt to find the rule of use of multiple techniques
(Absolute residual, MMRE)
Data set

- Take 77 observations (module) from Medical records database system
  - Each observation has 26 independent variables
    - Number of entities, number of attributes, number of transaction,..
  - Training set : Validation set = 52 : 25 (2 : 1)
- Use two group of training set TS1_{26}, TS2_{26}
  - One-off sampling can lead to misleading result

Goal

- Compare the accuracy of different prediction techniques
- Consider the effectiveness of multiple techniques
Least squares regression (LSR)

There's considerable inter-item correlation between the collected variables $\Rightarrow$ Collinearity

<table>
<thead>
<tr>
<th>Dependent variable(1)</th>
<th>Independent variables(25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effort</td>
<td>#entity, #transaction, #attribute,..</td>
</tr>
</tbody>
</table>

The number of attribute and effort have strong relationship

TS1: Effort = 1.147(ATRIBS)
TS2: Effort = 1.160(ATRIBS)

Very small intercepts $\Rightarrow$ ignore
Case study (3/8)

- Case-based reasoning (CBR)
  - Case-based reasoning (CBR)
    - Use ANGEL tool*
    - Increasing the number of analogies with the size of the case base is an effective strategy
    - Using the entire feature set even though there is evidence to suggest this is less than optimal
      - Searching ideal number of feature is very difficult in actuality

* ANGEL (ANaloGy softwarE Tool)
  - Developed by Bournemouth University in 1997
  - Software estimation tool by CBR
Comparison of the accuracy

<table>
<thead>
<tr>
<th>Technique</th>
<th>Sum of absolute residuals</th>
<th>Median absolute residual</th>
<th>Range</th>
<th>MMRE (%)</th>
<th>Bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A : TS1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert judgment</td>
<td>295.90</td>
<td>5.65</td>
<td>103.00</td>
<td>74.9</td>
<td>-26</td>
</tr>
<tr>
<td>LSR</td>
<td>241.26</td>
<td>4.92</td>
<td>79.27</td>
<td>45.5</td>
<td>-9</td>
</tr>
<tr>
<td>CBR</td>
<td>253.66</td>
<td>3.89</td>
<td>89.10</td>
<td>49.2</td>
<td>-6</td>
</tr>
<tr>
<td><strong>Panel B : TS2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert judgment</td>
<td>410.10</td>
<td>6.35</td>
<td>94.90</td>
<td>87.5</td>
<td>-20</td>
</tr>
<tr>
<td>LSR</td>
<td>223.32</td>
<td>5.55</td>
<td>40.50</td>
<td>43.2</td>
<td>-11</td>
</tr>
<tr>
<td>CBR</td>
<td>395.57</td>
<td>8.12</td>
<td>54.37</td>
<td>58.8</td>
<td>-2</td>
</tr>
</tbody>
</table>

- Absolute residual (AR) : |actual value – predicted value|
- Range : distance between largest predicted value and smallest predicted value
- Bias : (mean of predicted values – actual value) * 100

- LSR > CBR > Expert judgment
Comparison of the accuracy (cont’d)

- Frequencies of technique performing best
  - Best: minimum absolute residual for a particular module

<table>
<thead>
<tr>
<th>Prediction technique</th>
<th>TS1(26)</th>
<th>TS2(26)</th>
<th>Combined(52)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>LSR</td>
<td>9</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>CBR</td>
<td>11</td>
<td>8</td>
<td>19</td>
</tr>
</tbody>
</table>

- LSR is most effective
- But, only 40% ($\frac{21}{52}$) can find relative accuracy result

⇒ Need more effective technique!
Case study(6/8)

- Relationship between prediction techniques
  - Histogram of % prediction errors of a same module
  - Each technique tend to behave independently

⇒ Small possibility of multiple technique
Relationship between prediction techniques (cont’d)

- Covariance between absolute errors from prediction techniques

<table>
<thead>
<tr>
<th>Prediction technique</th>
<th>Expert Judgment</th>
<th>Least squares regression</th>
<th>Case-based reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert</td>
<td>TS1</td>
<td>TS2</td>
</tr>
<tr>
<td>Expert</td>
<td>0.612</td>
<td>0.745</td>
<td></td>
</tr>
<tr>
<td>LSR</td>
<td>0.100</td>
<td>-0.108</td>
<td>0.125</td>
</tr>
<tr>
<td>CBR</td>
<td>0.028</td>
<td>0.009</td>
<td>0.012</td>
</tr>
</tbody>
</table>

⇒ There’s possibility of complementarity among multiple prediction
Case study

Comparison of the accuracy to find the rule of multiple techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Panel A: TS1</th>
<th>Panel B: TS2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum of absolute error</td>
<td>Median of absolute error</td>
</tr>
<tr>
<td>LSR</td>
<td>241.26</td>
<td>4.92</td>
</tr>
<tr>
<td>Average of techniques</td>
<td>214.45</td>
<td>3.25</td>
</tr>
<tr>
<td>Using Rule Induction</td>
<td>224.66</td>
<td>5.15</td>
</tr>
<tr>
<td>Theoretical optimum</td>
<td>169.63</td>
<td>1.74</td>
</tr>
</tbody>
</table>

Ideal result: we always correctly pick the best technique.

⇒ Can’t find any rule of multiple prediction.
Conclusion

❖ Contribution
  ▪ There’s potential benefit in using more than one technique
    • Less or negative covariance between prediction techniques

❖ Further work
  ▪ Still need to find the reliable means of knowing a priori which technique to use