Mining Complex Feature Correlations from Software Product Line Configurations

Bo Zhang
University of Kaiserslautern
Kaiserslautern, Germany

Martin Becker
Fraunhofer IESE
Kaiserslautern, Germany
Agenda

• Motivation
• State-of-the-Art
• Our Approach
• Conclusion
Motivation

- **SPL Configuration Challenges** [PB+12] [DS 08]
  - Increasing number of features and feature values with their correlations
  - Misconfiguration due to missing/implicit feature correlations

- **Our Solution**
  - Deriving feature correlations from existing product configurations
  - Using the correlations as prediction knowledge for new product configuration

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td></td>
<td></td>
<td><strong>defined</strong></td>
<td><strong>defined</strong></td>
</tr>
<tr>
<td>F2</td>
<td>30</td>
<td>15</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>F3</td>
<td>EUR</td>
<td>USD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P5 defined

\[ F1 = \text{defined} \rightarrow F2 = 11 \]
State-of-the-Art: Reverse Engineering Variability

- Czarnecki et al. introduced Probabilistic FM with hard and soft feature constraints [CW 07] [CSW 08]

<table>
<thead>
<tr>
<th>Tid</th>
<th>Clock</th>
<th>Alarm</th>
<th>Digital</th>
<th>24hr</th>
<th>Analog</th>
<th>Hands</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

- Lora-Michiels et al. identified pairwise feature constraints from binary configurations [LS+10]
- Only identified simple correlations, e.g., \( F_1 \rightarrow F_2 \)

How to extract complex feature correlations from existing product configurations?

e.g., \((F_1 = 10 \land F_2 = 0xFF) \rightarrow (F_3 = \text{True} \land F_4 = \text{"EUR"})\)
Agenda

- Motivation
- State-of-the-Art
- **Our Approach**
  - Demonstrated with an industrial example
- Conclusion
Approach Framework

- Using Data Mining techniques to identify association rules as feature correlations
Configuration Extraction

• Input
  – Configuration files (in XML), or
  – Preprocessor code (#defines)

• Output
  – Configuration Matrix

<Product name="P1">
  <Item name="F1"/>
  <Item name="F2">
    <Sel val="30"/>
  </Item>
  <Item name="F3">
    <Sel val="EUR"/>
  </Item>
</Product>

1 // Product P1
2 #define F1
3 #define F2 30
4 #define F3 "EUR"
5 6 //P1 Implementation
7 ...

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td></td>
<td></td>
<td>defined defined</td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>30</td>
<td>15</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>F3</td>
<td>EUR</td>
<td>USD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Binary feature
Industrial Example: Configuration Extraction

- Large-scale SPL in ES domain
- Configuration Matrix
  - 100 products
  - 486 features

Feature Distribution across All Products

# of Features in Each Product

- # Non-Binary Features (avg 86)
- # Binary Features (avg 39)

Number of Selected Features vs Configurations of 100 Products
Data Preparation

- **Value Assignment for Binary Features**
  - Assigning "defined" to selected features
  - Do absent features mean "undefined" or "excluded"?
    * Finding exclusive correlations, e.g., F1 ↔ F3

- **Removing features with abnormal values**

- **Value Discretization**
  - Equally grouping continuous values into k partitions

- **Finally 480 valid features across 100 products**
  - 50 features discretized
Correlation Mining Using Data Mining Techniques

- Calculation of Frequent Itemsets
  - E.g., \{F_1=\text{defined}, F_2<15\}
  - \text{Supp}(\Gamma) = \frac{\{|P \mid \Gamma \text{ selected in } P\}|}{|P|}
    - E.g., \text{Supp}(\{F_1=\text{defined}, F_2<15\}) = \frac{2}{4} = 0.5

- Calculation of Association Rules
  - E.g., \text{F_1=\text{defined}} \rightarrow \text{F_2<15}
  - \text{Supp}(Y:A \rightarrow C) = \frac{\{|P \mid A \land C \text{ satisfies } P\}|}{|P|}
    - = \text{Supp}(A \land C)
    - E.g., \text{Supp}(F_1=\text{defined} \rightarrow F_2<15) = \text{Supp}(\{F_1=\text{defined}, F_2<15\}) = 0.5
  - \text{Conf}(Y:A \rightarrow C) = \frac{|\{P \mid A \land C \text{ satisfies } P\}|}{|\{P \mid A \text{ satisfies } P\}|}
    - = \frac{\text{Supp}(A \land C)}{\text{Supp}(A)}
    - E.g., \text{Conf}(F_1=\text{defined} \rightarrow F_2<15) = \frac{2}{3} = 0.66
Calculation of Frequent Itemsets

• Apriori Algorithm [AI+93]
  – Implemented in the Orange Tool [OT]

• Result of Industrial Example
  – 330 features * 100 products
  – 7581 frequent itemsets with min. support of 0.8
    • Avg. Support = 0.818, Avg. # Features = 5.43
Calculation of Association Rules

- Result of Industrial Example (330 features * 100 products)
  - 459,388 rules with min. support of 0.8 and min confidence of 0.9
  - Involving 11 binary features and 11 non-binary features

<table>
<thead>
<tr>
<th>Support</th>
<th>Confidence</th>
<th># Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.988</td>
<td>35601</td>
</tr>
<tr>
<td>0.8</td>
<td>1.000</td>
<td>31557</td>
</tr>
<tr>
<td>0.8</td>
<td>0.941</td>
<td>24873</td>
</tr>
<tr>
<td>0.8</td>
<td>0.976</td>
<td>24462</td>
</tr>
<tr>
<td>0.8</td>
<td>0.930</td>
<td>23685</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequent</th>
<th># Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>38766</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>34507</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>32105</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>31388</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>26760</td>
</tr>
</tbody>
</table>

**Confidence**

Avg = 0.96

**Support**

Avg = 0.81

**Size of Consequent**

Avg = 3.34

**Size of Antecedent**

Avg = 3.76
Correlation Pruning

- Reduce the number of association rules that do not offer any predictive advantage
- pruning a Sub-Rule that has a equal or smaller confidence than its parent rule
  - Parent rule: subset antecedent and superset consequent
    - \( A_p \rightarrow C_p = A_p \rightarrow (C_s \land C') = \neg A_p \lor (C_s \land C') \Rightarrow \neg A_p \lor C_s \Rightarrow \neg A_p \lor \neg A' \lor C_s \)
  - Sub-rule: superset antecedent and subset consequent
    - \( A_s \rightarrow C_s = (A_p \land A') \rightarrow C_s = \neg (A_p \land A') \lor C_s = \neg A_p \lor \neg A' \lor C_s \)
    - Hence, \( A_p \rightarrow C_p \Rightarrow A_s \rightarrow C_s \)

- Result
  - totally pruned 455022 and left 4366(1%) rules
  - Involving 11 binary features and 11 non-binary features
Distribution of 4366 Remaining Rules

### Distribution of Rules

<table>
<thead>
<tr>
<th>Support</th>
<th>Confidence</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.800</td>
<td>0.941</td>
<td>244</td>
</tr>
<tr>
<td>0.800</td>
<td>0.930</td>
<td>239</td>
</tr>
<tr>
<td>0.800</td>
<td>0.920</td>
<td>192</td>
</tr>
<tr>
<td>0.800</td>
<td>0.952</td>
<td>175</td>
</tr>
<tr>
<td>0.810</td>
<td>0.942</td>
<td>161</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Antecedent and Consequent Distribution

<table>
<thead>
<tr>
<th>Antecedent</th>
<th>Consequent</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6</td>
<td>507</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>455</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>419</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>409</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>400</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Comparison: before vs. after Pruning


Avg = 0.96  Avg = 0.81

Avg = 3.34  Avg = 3.76

Avg = 5.63  Avg = 2.33
Agenda

• Motivation
• State-of-the-Art
• Our Approach
• Conclusion
Conclusion

- Feature correlation mining framework
  - using data mining techniques
- Analysis of an industrial SPL example
  - 100 product configurations with 480 features
  - Finally got 4366 association rules after pruning

- Future work
  - Correlation validation with domain knowledge
  - Improving scalability of our correlation mining approach
References


