Cooperative Data Analysis in Supply Chains Using Selective Information Disclosure

JÖRG LÄSSIG¹ AND MICHAEL HAHSLER²

¹University of Applied Sciences Zittau/Görlitz, Germany
²Southern Methodist University, Dallas, Texas

INFORMS Computing Society Conference
Richmond, VA
January 2015
Global supply chain with many suppliers

Products become more complex

Finding defects requires access to data for analysis

Effective strategies for cooperative data analysis using selective data disclosure.
Background

Privacy preserving data mining
- Protect private information (age, income, etc.)
- Aim: Statistical data analysis on the aggregate

Companies participating in a supply chain
- Incentive to share (mostly logistics) information
  (Huang, Lau & Mak 2003, Subramani 2004)
- Competition can hinder information sharing (Li 2002, Frohlich 2002)
- Information protection goals are different than for companies
- Root cause analysis (RCA) needs not just logistics information
  - Details about a proprietary production process
  - Change of a third party supplier
  - Large volume of very detailed data

→ Trade-off: Minimize necessary information exchange
Supply Chain

- A directed acyclic graph $G = (V, E)$
- Participants $V$
- Material and information flows $E$

Vertically partitioned data set $\mathcal{T}$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$t_{1,1}$</th>
<th>$t_{1,2}$</th>
<th>...</th>
<th>$t_{1,m(v_1)}$</th>
<th>$t_{2,1}$</th>
<th>...</th>
<th>$t_{2,m(v_2)}$</th>
<th>...</th>
<th>$t_{12,1}$</th>
<th>...</th>
<th>$t_{12,m(v_{12})}$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m 1</td>
<td>s 3</td>
<td>...</td>
<td>m 2</td>
<td>s 4</td>
<td>...</td>
<td>e 4</td>
<td>...</td>
<td>e 3</td>
<td>...</td>
<td>m 6</td>
<td>pass</td>
</tr>
<tr>
<td>2</td>
<td>m 3</td>
<td>s 9</td>
<td>...</td>
<td>m 2</td>
<td>s 1</td>
<td>...</td>
<td>e 2</td>
<td>...</td>
<td>e 7</td>
<td>...</td>
<td>m 1</td>
<td>fail</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$</td>
<td>K</td>
<td>$</td>
<td>m 1</td>
<td>s 4</td>
<td>...</td>
<td>m 1</td>
<td>s 3</td>
<td>...</td>
<td>e 2</td>
<td>...</td>
<td>e 3</td>
<td>...</td>
</tr>
</tbody>
</table>

*Note.* m ... machine, s ... supplier, e ... employee.
Protocols for Optimized Information Disclosure

**Trivial case**
- Single party or full disclosure

**Direct case**
- Scenario 1: Known supplier
- Scenario 2: Supplier not known
- Scenario 3: Interaction between suppliers

**Remote case**
- Propagate the class information
- Recursive application of Direct case
Direct Case

**Scenario 1:** Known supplier
- Supplier $v$ receives class information for analysis.
- Addresses problems and/or reports results to $s$.

**Scenario 2:** Supplier not known
- All direct suppliers receive class information for analysis.
- One supplier finds a strong association addresses problem and notifies $s$.

**Scenario 3:** Interaction between suppliers
- Like scenario 2, but several supplier find (weaker) associations.
- Further analysis can be coordinated by $s$. 
Analyzing Dependencies

Many methods are available for root cause analysis

Statistical analysis
- Contingency tables and Chi-square test
- (rank) correlation

Data mining (Tan, Steinbach & Kumar, 2006)
- Classification model
  - Decision trees, logistic regression, SVM, etc.
  - Variable importance
- Association analysis
  - Association rule mining
## Association Rules

- Create transactions

### Table: \( T_{v1} \) and Class \( c \)

| \( k \) | \( t_{1,1} \) | \( t_{1,2} \) | ... | \( t_{1,m(v_1)} \) | \( t_{2,1} \) | ... | \( t_{2,m(v_2)} \) | ... | \( t_{12,1} \) | ... | \( t_{12,m(v_{12})} \) | \( c \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | m1 | s3 | ... | m2 | s4 | ... | e4 | ... | e3 | ... | m6 | pass |
| 2 | m3 | s9 | ... | m2 | s1 | ... | e2 | ... | e7 | ... | m1 | fail |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| \( |K| \) | m1 | s4 | ... | m1 | s3 | ... | e2 | ... | e3 | ... | m4 | pass |

*Note: m ... machine, s ... supplier, e ... employee.*

\( T_{v1} \bowtie c \)

<table>
<thead>
<tr>
<th>( t_{1,1} = m1 )</th>
<th>( t_{1,1} = m2 )</th>
<th>( t_{1,1} = m3 )</th>
<th>( t_{1,2} = s1 )</th>
<th>...</th>
<th>( t_{1,m(v_1)} = m1 )</th>
<th>( c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( t_{</td>
<td>K</td>
<td>} )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\( X \rightarrow \text{class} \)
Association Rules

Set of items: \( I = \{ t_{1,1} = m_1, t_{1,2} = m_2, \ldots, t_{1,m(v_1)} = m_1 \} \)

Left hand side of rule: \( X \subseteq I \)

Rule: \( X \rightarrow c \)

Support: \( \text{sup}(X \rightarrow c) = \frac{|\{k \in \{1,2,\ldots,|K|\} : X \cup c \subseteq t_k\}|}{|K|} > \sigma \)

Confidence: \( \text{conf}(X \rightarrow c) = \frac{\text{sup}(X \cup c)}{\text{sup}(X)} > \gamma \)
Association Rules

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>X̄</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>ċ</td>
<td>3000</td>
<td>400000</td>
</tr>
</tbody>
</table>

One-sided Fisher's exact test to measure the strength of rules (Hahsler & Hornik, 2007). Accept associations with a $p$-value $< \alpha$

- Correction for multiple comparisons (Miller, 1981)

Bonferroni Correction:

$$\alpha = \frac{\alpha^*}{m}$$

- Number of shared tuples $n=|\Gamma|$ represent a sample.

Upper limit on sample size based on Chernoff bounds (Zaki et al., 1997):

$$n = \frac{-2\ln(c)}{\tau \varepsilon^2}$$

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ε</td>
<td>\text{error rate}</td>
</tr>
<tr>
<td>1-c</td>
<td>\text{confidence level}</td>
</tr>
<tr>
<td>τ</td>
<td>\text{support}</td>
</tr>
</tbody>
</table>
Simulation Study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of suppliers</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>Features per supplier</td>
<td>$m(v)$</td>
<td>[5, 20]</td>
</tr>
<tr>
<td>Values per feature</td>
<td>$\text{range}(t_{v,i})$</td>
<td>[2, 10]</td>
</tr>
<tr>
<td>Base defect rate</td>
<td>$e_{\text{base}}$</td>
<td>1%</td>
</tr>
<tr>
<td>Feature defect rate</td>
<td>$e_{\text{feature}}$</td>
<td>100% or 5%</td>
</tr>
<tr>
<td>Number of shared tuples</td>
<td>$</td>
<td>\Gamma</td>
</tr>
<tr>
<td>Number of simulation runs</td>
<td>$n$</td>
<td>100</td>
</tr>
</tbody>
</table>

Chernoff bounds give 240,000 at 1% support and confidence and accuracy level of 95%

Scenario 1: Known supplier  
- Simple case of Scenario 2

Scenario 2: Supplier not known

Scenario 3: Interaction between 2 suppliers  
- Like Scenario 2, but several supplier find (weak) associations.  
- Further analysis can be coordinated by $s$. 
Amount of Shared Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of suppliers</td>
<td>$</td>
<td>V</td>
</tr>
<tr>
<td>Features per supplier</td>
<td>$m(v)$</td>
<td>[5, 20]</td>
</tr>
<tr>
<td>Values per feature</td>
<td>$\text{range}(t_{v,i})$</td>
<td>[2, 10]</td>
</tr>
<tr>
<td>Base defect rate</td>
<td>$e_{\text{base}}$</td>
<td>1%</td>
</tr>
<tr>
<td>Feature defect rate</td>
<td>$e_{\text{feature}}$</td>
<td>100% or 5%</td>
</tr>
<tr>
<td>Number of shared tuples</td>
<td>$</td>
<td>\Gamma</td>
</tr>
<tr>
<td>Number of simulation runs</td>
<td>$n$</td>
<td>100</td>
</tr>
</tbody>
</table>

Avg: 744 unique features values

Fixed at $|\Gamma| \times |V|$
Scenario 2: Supplier not known

Finding less frequent errors takes more data.

Selective disclosure is as effective as complete disclosure.

Selective disclosure incorrectly reports more features due to undercorrection.
Scenario 3: Interaction between 2 suppliers

Same as for Scenario 2:

Finding less frequent errors takes more data.

Selective disclosure is as effective as complete disclosure.

Selective disclosure incorrectly reports more features due to undercorrection.
Deployment

- Easy to use plug-in for RapidMiner.
- Central coordination web service to model supply chain.
- Secure communication directly between participants.
- Participants have full control over what information is shared.

https://rapidminer.com/
Conclusion

Many modern products are complicated and error-prone.

Data to perform root cause analysis is often not shared in supply chains.

Selective information disclosure:
  – Addresses need to perform distributed data analysis
  – Minimizes the amount of data to be exposed
  – Can be automated such that participants do not need to have in-depth data analysis capabilities
  – Initial experiments suggest that it can be effective
Thank you for your attention!

Michael Hahsler
mhahsler@lyle.smu.edu