Finding Statistically Significant Interactions between Continuous Features

Mahito Sugiyama (National Institute of Informatics)
Karsten Borgwardt (ETH Zürich)
Our Proposal: C-Tarone

- Find all feature interactions that are significantly associated with class labels from multivariate data with controlling the FWER

<table>
<thead>
<tr>
<th>ID1</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>...</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.96</td>
<td>-3.03</td>
<td>3.38</td>
<td>2.57</td>
<td>-6.06</td>
<td>...</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>-1.80</td>
<td>4.45</td>
<td>-4.35</td>
<td>0.82</td>
<td>8.90</td>
<td>...</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-3.29</td>
<td>1.39</td>
<td>-4.44</td>
<td>-0.77</td>
<td>2.78</td>
<td>...</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>-0.53</td>
<td>-1.96</td>
<td>-3.43</td>
<td>-4.42</td>
<td>-3.92</td>
<td>...</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Our Proposal: *C-Tarone*

- Find all feature interactions that are **significantly associated** with class labels from multivariate data **with controlling the FWER**

**Input:**

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>...</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>–0.96</td>
<td>–3.03</td>
<td>3.38</td>
<td>2.57</td>
<td>–6.06</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>ID2</td>
<td>–1.80</td>
<td>4.45</td>
<td>–4.35</td>
<td>0.82</td>
<td>8.90</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>ID3</td>
<td>–3.29</td>
<td>1.39</td>
<td>–4.44</td>
<td>–0.77</td>
<td>2.78</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>ID4</td>
<td>–0.53</td>
<td>–1.96</td>
<td>–3.43</td>
<td>–4.42</td>
<td>–3.92</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Output:**

{F1}, {F3},
{F2, F5},
{F2, F5, F6}, ...

---

1/14
Existing Method: Significant Pattern Mining

- So far only binary (or discrete) data can be used
  → Results obtained by SPM via binarization can be uninformative!

<table>
<thead>
<tr>
<th>ID</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID4</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Output:

- {F1}, {F3},
- {F2, F5},
- {F2, F5, F6}, ...
We solve:

1. How to assess the significance for a **multiplicative interaction of continuous features**?

2. How to perform **multiple testing correction**?
   - How to control the **FWER** (family-wise error rate), the probability to detect one or more false positives?

3. How to manage **combinatorial explosion** of the candidate space?
   - The number of possible interactions is $2^d$ for $d$ features
Problem Formulation

• Define $X_\mathcal{F}$ as the binary random variable of joint occurrence for a feature combination $\mathcal{F} = \{F_i, F_{i+1}, \ldots, F_{i+k}\}$
  - $X_\mathcal{F} = 1$ if $\mathcal{F}$ “occurs”, $X_\mathcal{F} = 0$ otherwise

• Let $Y$ be an output binary variable

• **Our task:** Test the null hypothesis $X_\mathcal{F} \perp \perp Y$ for all $\mathcal{F} \in 2^V$
  - Testing statistical independence between $X_\mathcal{F}$ and $Y$

• We need to estimate the probability $\Pr(X_\mathcal{F})$ from data
Copula Support [Tatti, 2013] for $\Pr(X_F = 1)$

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>$R(F1)$</th>
<th>$R(F2)$</th>
<th>$R(F3)$</th>
<th>$\pi(F1)$</th>
<th>$\pi(F2)$</th>
<th>$\pi(F3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>-0.96</td>
<td>-3.03</td>
<td>3.38</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0.67</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-1.80</td>
<td>4.45</td>
<td>-4.35</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0.34</td>
<td>1.00</td>
<td>0.34</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-3.29</td>
<td>1.39</td>
<td>-4.44</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-0.53</td>
<td>-1.96</td>
<td>-3.43</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1.00</td>
<td>0.34</td>
<td>0.67</td>
</tr>
</tbody>
</table>

$$0.083 = \Pr(X_{\{F1,F2,F3\} = 1}) = \eta(\{F1,F2,F3\})$$
## Contingency Tables

<table>
<thead>
<tr>
<th></th>
<th>( X_\mathcal{F} = 1 )</th>
<th>( X_\mathcal{F} = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expected (under null) for ( p_E )</strong></td>
<td>( \eta(\mathcal{F}) r_1 )</td>
<td>( r_1 - \eta(\mathcal{F}) r_1 )</td>
<td>( r_1 )</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y = 0 )</td>
<td></td>
<td>( \eta(\mathcal{F}) r_0 )</td>
<td>( r_0 )</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>( \eta(\mathcal{F}) )</td>
<td>( 1 - \eta(\mathcal{F}) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( X_\mathcal{F} = 1 )</th>
<th>( X_\mathcal{F} = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed for ( p_O )</strong></td>
<td>( \eta(\mathcal{F}, Y = 1) )</td>
<td>( r_1 - \eta(\mathcal{F}, Y = 1) )</td>
<td>( r_1 )</td>
</tr>
<tr>
<td>( Y = 1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Y = 0 )</td>
<td>( \eta(\mathcal{F}, Y = 0) )</td>
<td>( r_0 - \eta(\mathcal{F}, Y = 0) )</td>
<td>( r_0 )</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>( \eta(\mathcal{F}) )</td>
<td>( 1 - \eta(\mathcal{F}) )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>
Significance Test

- The independence $X \independent F \perp \perp Y$ is translated into the condition:
  
  $H_0 : D_{KL}(p_O, p_E) = 0, \quad H_1 : D_{KL}(p_O, p_E) \neq 0$
  
  - $p_E$ and $p_O$ are vectorized contingency tables:
    
    $p_E = (\eta(F)r_1, \eta(F)r_0, r_1 - \eta(F)r_1, r_0 - \eta(F)r_0)$
    
    $p_O = (\eta(F, Y=1), \eta(F, Y=0), r_1 - \eta(F, Y=1), r_0 - \eta(F, Y=0))$

- We apply **G-test**: the statistic $\lambda = 2ND_{KL}(p_O, p_E)$ follows the $\chi^2$-distribution with the d.f. 1
Multiple Testing Correction

- The FWER should be controlled
  - Probability that at least one feature combination is a false positive
  - If we naively test all combinations, $a2^d$ false positives could occur!!

- We use Tarone’s testability trick, which requires the minimum achievable $p$-value $\psi(\mathcal{F})$ for $\mathcal{F}$

- **Theorem** (tight upper bound of KL divergence):

  $$D_{KL}(\mathbf{p}, \mathbf{p}_E) < a \log \frac{1}{b} + (b - a) \log \frac{b - a}{(1 - a)b} + (1 - b) \log \frac{1}{(1 - a)}$$

  - $\mathbf{p}_E = (ab, a(1 - b), (1 - a)b, (1 - a)(1 - b))$,
  - $\mathbf{p} \in \{ \mathbf{p} \in \mathcal{P} \mid p_1 + p_2 = a, p_1 + p_3 = b \}$
Tarone’s Testability Trick

\[ \mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots, \mathcal{F}_{m-1}, \mathcal{F}_m, \mathcal{F}_{m+1}, \ldots, \mathcal{F}_{2d} \quad (\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1})) \]
Tarone’s Testability Trick

\[ m \psi(F_m) < \alpha \text{ and } (m + 1)\psi(F_{m+1}) \geq \alpha \]

\[ F_1, F_2, F_3, ..., F_{m-1}, F_m, F_{m+1}, ..., F_{2d} \quad (\psi(F_i) \leq \psi(F_{i+1})) \]
Tarone’s Testability Trick

\[ m \psi(F_m) < \alpha \quad \text{and} \quad (m+1)\psi(F_{m+1}) \geq \alpha \]

\[ F_1, F_2, F_3, ..., F_{m-1}, F_m, F_{m+1}, ..., F_{2d} \quad (\psi(F_i) \leq \psi(F_{i+1})) \]

Testable combinations

Untestable combinations  \rightarrow Prune without testing

\[ F_i \text{ is significant if: } p\text{-value}(F_i) < \frac{\alpha}{m} \]

Correction factor
Enumeration Algorithm Based on Apriori

F1, F2, F3, Fd
Fd–1, Fd
F1, F2
F1, F2, F3
Fd
Enumeration Algorithm Based on Apriori

Threshold for $\eta$

Smaller $\eta$ $\rightarrow$ Larger $\psi$
Enumeration Algorithm Based on Apriori

Smaller $\eta \rightarrow$ Larger $\psi$
Enumeration Algorithm Based on Apriori

F1, F2, F3, Fd
Fd–1, Fd, F1, F2

Smaller $\eta$ $\rightarrow$ Larger $\psi$
Enumeration Algorithm Based on Apriori

Testable combinations $F_1, F_2, F_3, F_d$

Smaller $\eta$ → Larger $\psi$
Experimental Results on Synthetic Data

- **F-measure**
  - Number of data points: $10^3$ to $10^5$
  - Number of features: $10$ to $100$

**Graphs**

- **F-measure** vs. **Number of data points**
- **F-measure** vs. **Number of features**

**Lines**

- C-Tarone
- Binarization

**Legend**

- C-Tarone
- Binarization

11/14
Experimental Results on Synthetic Data

![Graph showing running time vs. number of data points and features]

- **Running time (sec.)**
  - **Number of data points**
    - $10^3$, $10^4$, $10^5$
  - **Number of features**
    - $20$, $40$, $60$, $80$, $100$

- **C-Tarone**
- **Binarization**
Experimental Results on Real Data

<table>
<thead>
<tr>
<th></th>
<th>C-Tarone (proposed)</th>
<th>Binarization</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faults</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ijcnn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>magic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>segment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F-measure

Running time (sec.)
Experimental Results on Real Data

![Graph showing F-measure and Running time (sec.) for different datasets: transfusion, waveform, wdbc, wine, yacht.]

- **C-Tarone (proposed)**
- **Binarization**
Conclusion

- We have proposed **C-Tarone**, a solution to the open problem of finding *all* multiplicative interactions between *continuous* features significantly associated with an output variable
  - Significance is rigorously controlled for multiple testing

- Our work opens the door to many applications of searching significant feature combinations, in which the data is not adequately described by binary features