Identifying Consistent Statements about Numerical Data with Dispersion-Corrected Subgroup Discovery

Mario Boley, Bryan Goldsmith, Luca Ghiringhelli, Jilles Vreeken
mboley@mpi-inf.mpg.de

Max Planck Institute for Informatics and Saarland University
Fritz Haber Institute of the Max Planck Society
Subgroup discovery (with binary target)

**Given**
Sample $S \subseteq P$
Target variable $y: P \rightarrow \{\oplus, \ominus\}$
Description variables $x_j: P \rightarrow X_j$

[Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008]
Subgroup discovery (with binary target)

**Given**
Sample $S \subseteq P$
Target variable $y: P \rightarrow \{\oplus, \ominus\}$
Description variables $x_j: P \rightarrow X_j$

**Define**
Selection language $\mathcal{L}_x \subseteq \{\bot, T\}^P$
($\sigma \in \mathcal{L}_x$ defines $\text{ext}(\sigma) = \{i \in S: \sigma(i) = T\} \subseteq S$)

[ Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008 ]
Subgroup discovery (with binary target)

Given
Sample $S \subseteq P$
Target variable $y: P \rightarrow \{\oplus, \ominus\}$
Description variables $x_j: P \rightarrow X_j$

Define
Selection language $\mathcal{L}_x \subseteq \{\bot, T\}^P$
($\sigma \in \mathcal{L}_x$ defines $\text{ext}(\sigma) = \{i \in S: \sigma(i) = T\} \subseteq S$)

Optimize
$f(Q) = \text{cov}(Q) \gamma \text{eff}(Q)_+$
with
- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$ coverage
- $\text{eff}(Q) = \tilde{y}(Q) - \tilde{y}(S)$ effect
- $\tilde{y}(Q) = \{i \in Q: y(i) = \oplus\}/|Q|$ pos. prob.

[Klösgen, 1996; Wrobel, 1997; Duivesteijn et al., 2008]
Better than global models in capturing *local* effects.
Metric target variables

**Given**

Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

**Define**

Selection language $\mathcal{L}_x \subseteq \{\perp, \top\}^P$
$(\sigma \in \mathcal{L}_x$ defines $\text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S)$

**Optimize**

$$f(Q) = \text{cov}(Q)y_{\text{eff}}(Q)_+$$

with

- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$ coverage
- $\text{eff}(Q) = \bar{y}(Q) - \bar{y}(S)$ effect
- $\bar{y}(Q)$ central tendency (mean, median,...)
Metric target variables

**Given**
Sample $S \subseteq P$
Target variable $y: P \to \mathbb{R}$
Description variables $x_j: P \to X_j$

**Define**
Selection language $\mathcal{L}_X \subseteq \{\bot, \top\}^P$
($\sigma \in \mathcal{L}_X$ defines $\text{ext}(\sigma) = \{i \in S: \sigma(i) = \top\} \subseteq S$)

**Optimize**
$$f(Q) = \text{cov}(Q)Y \text{eff}(Q)$$
with
- $Q = \text{ext}(\sigma)$
- $\text{cov}(Q) = |Q|/|S|$ coverage
- $\text{eff}(Q) = \bar{y}(Q) - \bar{y}(S)$ effect
- $\bar{y}(Q)$ central tendency (mean, median,...)
Dysfunctionality of pure coverage/effect approach

Dispersion

average error $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case $\hat{y}(Q) = \bar{y}(Q)$: $e(i) = (\hat{y}(Q) - y(i))^2$

case $\hat{y}(Q) = \text{med}(Q)$: $e(i) = |\hat{y}(Q) - y(i)|$

©2017, Mario Boley
Dysfunctionality of pure coverage/effect approach

Dispersion
average error $\overline{e}(Q) = \sum_{i \in Q} e(i) / |Q|$

case $\hat{y}(Q) = \overline{y}(Q)$: $e(i) = (\hat{y}(Q) - y(i))^2$

case $\hat{y}(Q) = \text{med}(Q)$: $e(i) = |\hat{y}(Q) - y(i)|$

Selection preference
any

\[ f(Q) = g(\text{cov}(Q), \text{eff}(Q)) \]

monotone in first argument favors groups that

- are not summarized well by $\hat{y}$
Dysfunctionality of pure coverage/effect approach

**Dispersion**

average error $\overline{e}(Q) = \sum_{i \in Q} \frac{e(i)}{|Q|}$

case $\tilde{y}(Q) = \bar{y}(Q)$: $e(i) = (\tilde{y}(Q) - y(i))^2$

case $\tilde{y}(Q) = \text{med}(Q)$: $e(i) = |\tilde{y}(Q) - y(i)|$

**Selection preference**

any

$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$

monotone in first argument favors groups that

- are not summarized well by $\tilde{y}$
- contain noise
Dysfunctionality of pure coverage/effect approach

**Dispersion**
average error $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

- case $\hat{y}(Q) = \bar{y}(Q)$: $e(i) = (\hat{y}(Q) - y(i))^2$
- case $\hat{y}(Q) = \text{med}(Q)$: $e(i) = |\hat{y}(Q) - y(i)|$

**Selection preference**
any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$

monotone in first argument favors groups that

- are not summarized well by $\hat{y}$
- contain noise
- are incoherent

\[ \sigma(i) \equiv a(i) \in [8,12] \land c_2(i) > \text{low} \land c_6(i) < \text{vhigh} \land r(i) > \text{vlow} \]
Dysfunctionality of pure coverage/effect approach

Dispersion
average error \( \bar{e}(Q) = \sum_{i \in Q} e(i)/|Q| \)

case \( \hat{y}(Q) = \bar{y}(Q) \):

\[ e(i) = (\hat{y}(Q) - y(i))^2 \]

case \( \hat{y}(Q) = \text{med}(Q) \):

\[ e(i) = |\hat{y}(Q) - y(i)| \]

Selection preference
any

\[ f(Q) = g(\text{cov}(Q), \text{eff}(Q)) \]

monotone in first argument favors groups that

- are not summarized well by \( \hat{y} \)
- contain noise
- are incoherent
Dysfunctionality of pure coverage/effect approach

Dispersion
average error $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$

case $\tilde{y}(Q) = \bar{y}(Q)$: $e(i) = (\tilde{y}(Q) - y(i))^2$

case $\tilde{y}(Q) = \text{med}(Q)$: $e(i) = |\tilde{y}(Q) - y(i)|$

Selection preference

any

$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$

monotone in first argument favors groups that

• are not summarized well by $\tilde{y}$
• contain noise
• are incoherent
Dysfunctionality of pure coverage/effect approach

Dispersion
average error $\bar{e}(Q) = \sum_{i \in Q} e(i)/|Q|$
case $\bar{y}(Q) = \bar{y}(Q)$: $e(i) = (\bar{y}(Q) - y(i))^2$
case $\bar{y}(Q) = \text{med}(Q)$: $e(i) = |\bar{y}(Q) - y(i)|$

Selection preference
any

$$f(Q) = g(\text{cov}(Q), \text{eff}(Q))$$
monotone in first argument favors groups that

- are not summarized well by $\bar{y}$
- contain noise
- are incoherent
- provide weak quantitative guarantees wrt $P$
In 10 of 25 datasets local error larger than global.
How to fix this?

Normalize effect ??

\[ f(Q) = g\left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right) \]

- ugly edge cases
- unclear how to optimize

\[ \tilde{y}_1 = \tilde{y}_2 \]

\[ \tilde{y}_2 - \epsilon_2 \]

\[ \tilde{y}_1 - \epsilon_1 \]

\[ \sigma_1(i) \equiv x_1(i) \geq 0.5 \]

\[ \sigma_2(i) \equiv (x_1(i) \geq 0.5) \land (x_2(i) = a) \]

[ Klösgen, 2002; Pieters et al., 2010 ]
How to fix this?

Normalize effect ??

$$f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right)$$

• ugly edge cases
• unclear how to optimize

Correct coverage

$$f(Q) = g \left( \text{cov}(Q) \left( \frac{\bar{e}(S) - \bar{e}(Q)}{\bar{e}(S)} \right), \text{eff}(Q) \right)$$
How to fix this?

Normalize effect ??

\[ f(Q) = g \left( \text{cov}(Q), \frac{\text{eff}(Q)}{\bar{e}(Q)} \right) \]

- ugly edge cases
- unclear how to optimize

Correct coverage

\[ f(Q) = g \left( \text{cov}(Q) \left( \frac{\bar{e}(S) - \bar{e}(Q)}{\bar{e}(S)} \right), \text{eff}(Q) \right) \]

\[ \text{dcc}(Q) = \left( \frac{|Q|}{|S|} - \frac{e(Q)}{e(S)} \right) \]

dispersion-corrected coverage

\[ \sigma_1(i) \equiv x_1(i) \geq 0.5 \]

\[ \sigma_2(i) \equiv (x_1(i) \geq 0.5) \land (x_2(i) = a) \]
How to fix this?

Normalize effect ??

\[ f(Q) = g\left( \text{cov}(Q), \frac{\text{eff}(Q)}{\overline{e}(Q)} \right) \]

- ugly edge cases
- unclear how to optimize

Correct coverage

\[ f(Q) = g\left( \text{cov}(Q), \frac{\overline{e}(S) - \overline{e}(Q)}{\overline{e}(S)} \right) + \text{eff}(Q) \]

\[ \text{dcc}(Q) = \left( \frac{|Q|}{|S|} - \frac{e(Q)}{e(S)} \right) + \text{dispersion-corrected coverage} \]

- well-behaved edge cases
- can be optimized w/o loosing performance (median case) !!
Dispersion-correction reduces error

\[
\tilde{y}(Q_0^*) \pm \overline{e}(Q_0^*) \quad \text{— classic} \quad f
\]

\[
\tilde{y}(Q_1^*) \pm \overline{e}(Q_1^*) \quad \text{— dispersion-corrected} \quad f
\]
Increases conservative mean estimates significantly
Branch-and-bound subgroup optimization

Branch

\[ r: \mathcal{L}_x \to 2^{\mathcal{L}_x} \]
\[ \varphi \in r(\sigma) \Rightarrow \sigma \preceq \varphi \Rightarrow \text{ext}(\sigma) \supseteq \text{ext}(\varphi) \]

Bound

\[ \hat{f}(\sigma) \geq \max\{f(\varphi) : \varphi \succeq \sigma\} \]

\[ \hat{f}(\sigma_5) \leq f(\sigma^*) \]

[Wrobel 1997; Grosskreutz et al. 2008; Boley and Grosskreutz 2009]
Branch-and-bound subgroup optimization

**Branch**

\[
\begin{align*}
\mathbf{r}: \mathcal{L}_x &\rightarrow 2^{\mathcal{L}_x} \\
\varphi \in \mathbf{r}(\sigma) &\Rightarrow \sigma \preceq \varphi \\
\Rightarrow \text{ext}(\sigma) &\supseteq \text{ext}(\varphi)
\end{align*}
\]

**Bound**

\[
\hat{f}(\sigma) = \max\{f(R): R \subseteq \text{ext}(\sigma)\} \\
\geq \max\{f(\varphi): \varphi \succeq \sigma\}
\]

tight optimistic estimator

\[
\hat{f}(\sigma_5) \leq f(\sigma^*)
\]

[Wrobel 1997; Grosskreutz et al. 2008; Boley and Grosskreutz 2009]
How to compute tight opt est in linear time?

**Given**

\[ f(Q) = g(\text{cov}(Q), \text{eff}(Q)) \]

\[ Q = \{y_1, \ldots, y_m\} \text{ st } y_1 \leq \cdots \leq y_m \]

[cf. Lemmerich et al 2016]
**Given**

\[ f(Q) = g(\text{cov}(Q), \text{eff}(Q)) \]
\[ Q = \{y_1, \ldots, y_m\} \text{ st } y_1 \leq \cdots \leq y_m \]

**Compute \( \hat{f} \) in time \( O(m) \)**

\[ \hat{f}(Q) = \max\{f(T_l): 1 \leq l \leq m\} \]
\[ T_l = \{y_{m-l+1}, \ldots, y_m\} \]

[cf. Lemmerich et al 2016]
Given

\[ f(Q) = g(\text{cov}(Q), \text{eff}(Q)) \]

\[ Q = \{y_1, \ldots, y_m\} \text{ st } y_1 \leq \ldots \leq y_m \]

Compute \( \hat{f} \) in time \( O(m) \)

\[ \hat{f}(Q) = \max\{f(T_l): 1 \leq l \leq m\} \]

\[ T_l = \{y_{m-l+1}, \ldots, y_m\} \]

Using incremental \( O(1) \) ops

\[ \text{cov}(T_l) = l/m \]

\[ \overline{y}(T_{l+1}) = \frac{l\overline{y}(T_l) + y_{m-l}}{l+1} \]

[cf. Lemmerich et al 2016]
How can we extend idea to dcc?

Given

\[ f(Q) = g(\text{dcc}(Q), \text{eff}(Q)) \]

\[ Q = \{y_1, \ldots, y_m\} \text{ st } y_1 \leq \cdots \leq y_m \]
Linear size Pareto front through median index

Given

\[ f(Q) = g(dcc(Q), \text{eff}(Q)) \]
\[ Q = \{y_1, \ldots, y_m\} \text{ s.t. } y_1 \leq \cdots \leq y_m \]

Compute \( \hat{f} \) in time \( O(m) \)

\[ \hat{f}(Q) = \max\{f(M_z): 1 \leq z \leq m\} \]
\[ M_z = M_z^{k_z} \]
\( k_z \) maximizing \( \text{dcc}(M_z^k) \)

\[ M_z^k = \left\{ y - \left\lfloor \frac{k}{2} \right\rfloor, \ldots, y_z, \ldots, y_z + \left\lfloor \frac{k}{2} \right\rfloor \right\} \]
Linear size Pareto front through median index

Given
\[ f(Q) = g(dcc(Q), \text{eff}(Q)) \]
\[ Q = \{ y_1, \ldots, y_m \} \text{ st } y_1 \leq \cdots \leq y_m \]

Compute \( \hat{f} \) in time \( O(m) \)
\[ \hat{f}(Q) = \max\{f(M_z): 1 \leq z \leq m\} \]
\[ M_z = M_z^{k_z^*} \]
\[ k_z^* \text{ maximizing } dcc(M_z^k) \]
\[ M_z^k = \left\{ y_{z - \lfloor k_z^* \rfloor}, \ldots, y_{z + \lceil k_z^* \rceil} \right\} \]

Using incremental \( O(1) \) ops
\[ \tilde{y}(M_z) = y_z \]
\[ dcc(M_z) \text{ ???} \]
Incremental computation of $k_Z^*$ in $O(1)$

Idea
Analyze functions

$h_z: k \mapsto \text{dcc}(M_Z^k)$

$\Delta h_z: k \mapsto (h_z(k) - h_z(k - 1))$
Incremental computation of $k_Z^*$ in $O(1)$

**Idea**

Analyze functions

$h_Z: k \mapsto \text{dcc}(M_Z^k)$

$\Delta h_Z: k \mapsto (h_Z(k) - h_Z(k - 1))$

**Observations**

$h_Z$ alternating concave, i.e.,

$\Delta h_Z(k + 1) + \Delta h_Z(k) \leq \Delta h_Z(k) + \Delta h_Z(k - 1)$
Incremental computation of $k^*_Z$ in $O(1)$

**Idea**
Analyze functions

$h_Z: k \mapsto \text{dcc}(M^k_Z)$

$\Delta h_Z: k \mapsto (h_Z(k) - h_Z(k - 1))$

**Observations**
$h_Z$ alternating concave, i.e.,

$\Delta h_Z(k + 1) + \Delta h_Z(k) \leq \Delta h_Z(k) + \Delta h_Z(k - 1)$
Incremental computation of $k^*_Z$ in $O(1)$

**Idea**
Analyze functions

$h_Z: k \mapsto \text{dcc}(M_Z^k)$

$\Delta h_Z: k \mapsto (h_Z(k) - h_Z(k - 1))$

**Observations**
$h_Z$ alternating concave, i.e.,

$\Delta h_Z(k + 1) + \Delta h_Z(k) \leq \Delta h_Z(k) + \Delta h_Z(k - 1)$

$\Delta h_{Z+1}$ and $\Delta h_Z$ are coupled, i.e.,

$\Delta h_{Z-1}(k) + \Delta Z_{-1}(k - 1) \leq \Delta h_Z(k - 2) + \Delta Z(k - 3)$
Incremental computation of $k^*_Z$ in $O(1)$

Idea
Analyze functions
\[ h_Z: k \mapsto \text{dcc}(M^k_Z) \]
\[ \Delta h_Z: k \mapsto (h_Z(k) - h_Z(k - 1)) \]

Observations
$h_Z$ alternating concave, i.e.,
\[ \Delta h_Z(k + 1) + \Delta h_Z(k) \leq \Delta h_Z(k) + \Delta h_Z(k - 1) \]
\[ \Delta h_{z+1} \text{ and } \Delta h_z \text{ are coupled, i.e.,} \]
\[ \Delta h_{z-1}(k) + \Delta h_{z-1}(k - 1) \leq \Delta h_Z(k - 2) + \Delta_Z(k - 3) \]

Conclusion
$k^*_Z \in \{k^*_{z+1} - 3, ..., k^*_{z+1} + 3\}$
-> incremental $O(1)$ computation
Gains of tight estimator over top-sequence-based
Conclusion

Summary

dispersion/error is an issue
dispersion-corrected coverage addresses it
can be optimized effectively (median case)

Directions

immediate: what about mean case?
bigger picture: complex models / multiple targets

[find resources at www.realkd.org]