Finding Statistically Significant Interactions between Continuous Features

Mahito Sugiyama (National Institute of Informatics), Karsten Borgwardt (ETH Zürich)

The 28th International Joint Conference on Artificial Intelligence (IJCAI-19), August 10–16, 2019

Our Proposal: *C-Tarone*

- Find all feature interactions that are significantly associated with class labels from multivariate data with controlling the FWER
  - Existing methods (significant pattern mining) work only for binary (or discrete) data [1]

Input:

	x					y
	F1	F2	F3	F4	F5	... Class
ID1	–0.96	–3.03	3.38	2.57	–6.06	... 0
ID2	–1.80	4.45	–4.35	0.82	8.90	... 1
ID3	–3.29	1.39	–4.44	–0.77	2.78	... 1
ID4	–0.53	–1.96	–3.43	–4.42	–3.92	... 0
⋮			⋮			⋮

→ **Output:** {F1}, {F3}, {F2, F5}, {F2, F5, F6}, ...

Significance Test for Feature Combination

- **Our task:** Test the null hypothesis $X_{\mathcal{F}} \perp\!\!\!\perp Y$ for *all* $\mathcal{F} \in 2^V$
  - $X_{\mathcal{F}}$: The binary random variable of joint occurrence for $\mathcal{F}$
- Copula Support [2] for $\Pr(X_{\mathcal{F}} = 1)$:

	F1	F2	F3		R(F1)	R(F2)	R(F3)		π(F1)	π(F2)	π(F3)
$\boldsymbol{x}_1$	–0.96	–3.03	3.38		2	0	3		0.67	0.00	1.00
$\boldsymbol{x}_2$	–1.80	4.45	–4.35	Rank →	1	3	1	Norm. →	0.34	1.00	0.34
$\boldsymbol{x}_3$	–3.29	1.39	–4.44		0	2	0		0.00	0.67	0.00
$\boldsymbol{x}_4$	–0.53	–1.96	–3.43		3	1	2		1.00	0.34	0.67

Prod. → 0.00, 0.11, 0.00, 0.22 → Sum / 4 → $0.083 = \Pr(X_{\{F1,F2,F3\}} = 1) = \eta(\{F1,F2,F3\})$

- The independence $X_{\mathcal{F}} \perp\!\!\!\perp Y$ is translated into the condition:
  $H_0 : D_{KL}(\boldsymbol{p}_O, \boldsymbol{p}_E) = 0, \quad H_1 : D_{KL}(\boldsymbol{p}_O, \boldsymbol{p}_E) \neq 0$
  - We apply G-test: $\lambda = 2ND_{KL}(\boldsymbol{p}_O, \boldsymbol{p}_E)$ follows χ^2-dist. with d.f. 1

Expected (under null) for $\boldsymbol{p}_E$	$X_{\mathcal{F}} = 1$	$X_{\mathcal{F}} = 0$	Total
$Y = 1$	$\eta(\mathcal{F})\, r_1$	$r_1 - \eta(\mathcal{F})\, r_1$	r_1
$Y = 0$	$\eta(\mathcal{F})\, r_0$	$r_0 - \eta(\mathcal{F})\, r_0$	r_0
Total	$\eta(\mathcal{F})$	$1 - \eta(\mathcal{F})$	1

Observed for $\boldsymbol{p}_O$	$X_{\mathcal{F}} = 1$	$X_{\mathcal{F}} = 0$	Total
$Y = 1$	$\eta(\mathcal{F}, Y = 1)$	$r_1 - \eta(\mathcal{F}, Y = 1)$	r_1
$Y = 0$	$\eta(\mathcal{F}, Y = 0)$	$r_0 - \eta(\mathcal{F}, Y = 0)$	r_0
Total	$\eta(\mathcal{F})$	$1 - \eta(\mathcal{F})$	1

Multiple Testing Correction

- The FWER should be controlled
  - Probability that at least one feature combination is a false positive
  - If we naïvely test all combinations, $\alpha 2^d$ false positives could occur!!
- We use Tarone's testability trick [3], which requires the minimum achievable p-value $\psi(\mathcal{F})$ for $\mathcal{F}$
- **Theorem** (tight upper bound of KL divergence):

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

  - $\boldsymbol{p}_E = (ab, a(1 - b), (1 - a)b, (1 - a)(1 - b))$
  - $\boldsymbol{p} \in \mathcal{P}(a, b) = \{ \boldsymbol{p} \in \mathcal{P} \mid p_1 + p_2 = a,\ p_1 + p_3 = b \}$

Tarone's Testability Trick

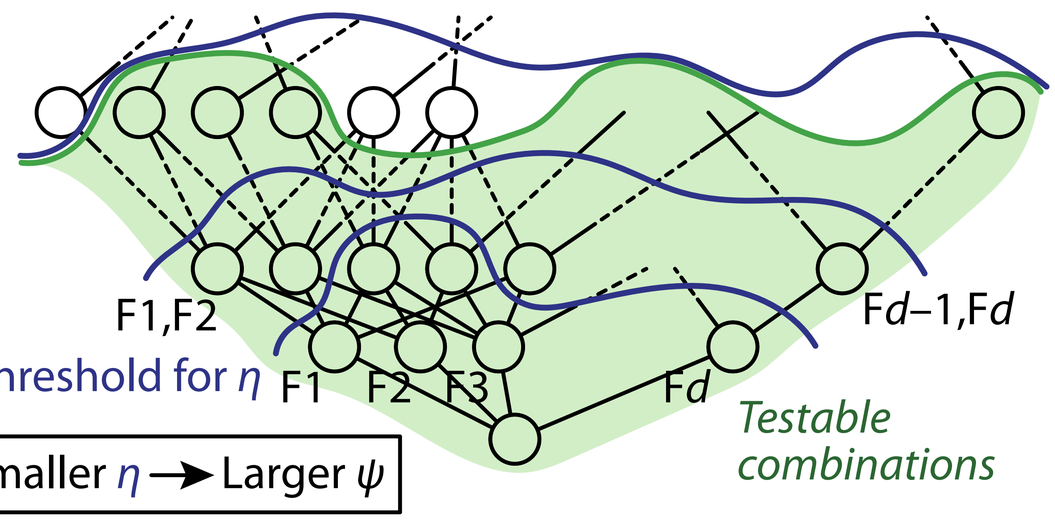
$m\,\psi(\mathcal{F}_m) < \alpha$ and $(m + 1)\psi(\mathcal{F}_{m+1}) \geq \alpha$

$\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots, \mathcal{F}_{m-1}, \mathcal{F}_m$ | $\mathcal{F}_{m+1}, \ldots, \mathcal{F}_{2^d}$ $\quad (\psi(\mathcal{F}_i) \leq \psi(\mathcal{F}_{i+1}))$

Testable combinations | Untestable combinations → Prune without testing

$\mathcal{F}_i$ is significant if: $p\text{-value}(\mathcal{F}_i) < \alpha / (m)$ — Correction factor

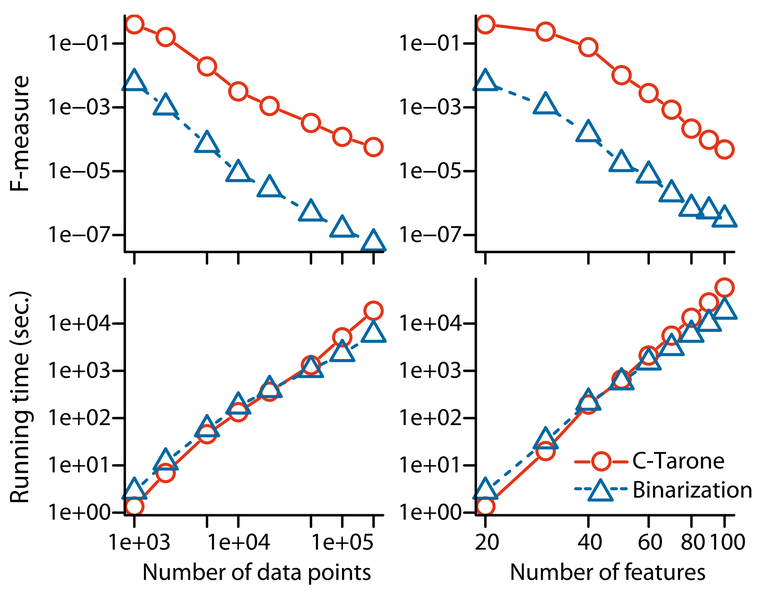
Enumeration Based on Apriori



Smaller η → Larger ψ

Experimental Results

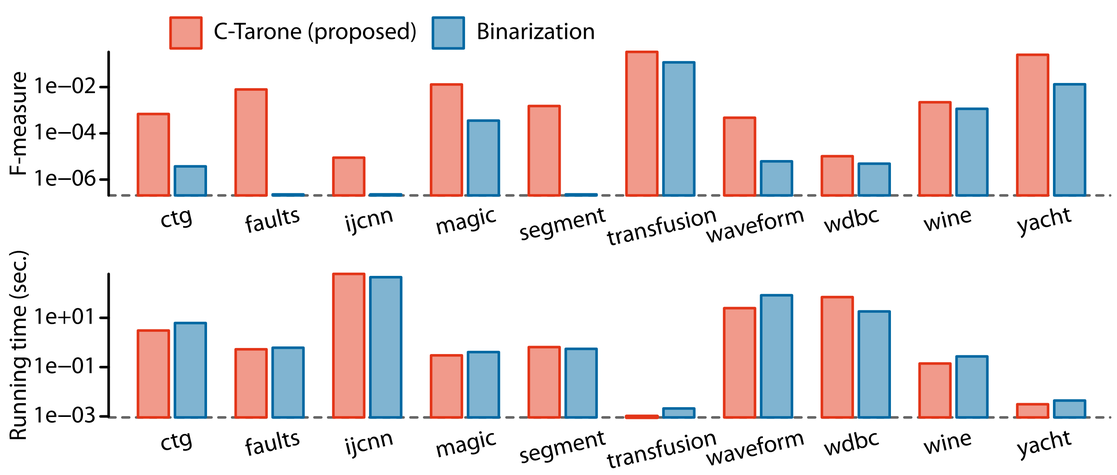
- Synthetic Data:



Statistics of real data

Data	N	d
ctg	942	22
faults	316	27
ijcnn	9,706	22
magic	13,376	10
segment	660	19
transfusion	356	4
waveform	3,314	21
wdbc	424	30
wine	3,198	11
yacht	308	6

- Real data:



Reference

- [1] Llinares-López, F., Sugiyama, M., Papaxanthos, L., Borgwardt, K.M.: Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing, *KDD 2015*
- [2] Tatti, N.: Itemsets for Real-Valued Datasets, *ICDM 2013*
- [3] Tarone, R.: A modified Bonferroni method for discrete data, *Biometrics*, 1990