DENSO IT LAB

AIST

ICML2019

Breaking Inter-Layer Co-Adaptation by Classifier Anonymization

Ikuro Sato[1]
Kohta Ishikawa[1]
Guoqing Liu[1]
Masayuki Tanaka[2]

[1] Denso IT Laboratory. Inc., Japan
[2] National Institute of Advanced Industrial Science and Technology, Japan

Summary first

About what? **Breaking co-adaptation** between feature extractor and classifier.

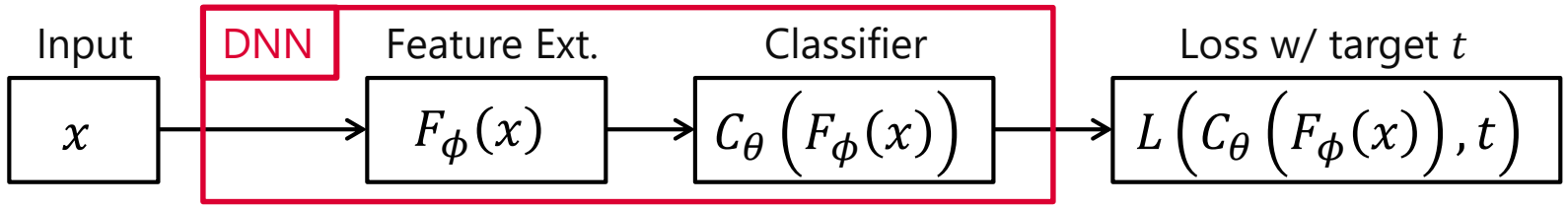
How? By **classifier anonymization** technique.

Theory? Proved: Features form simple **point-like distribution**.

In reality? Point-like property largely confirmed on real datasets.

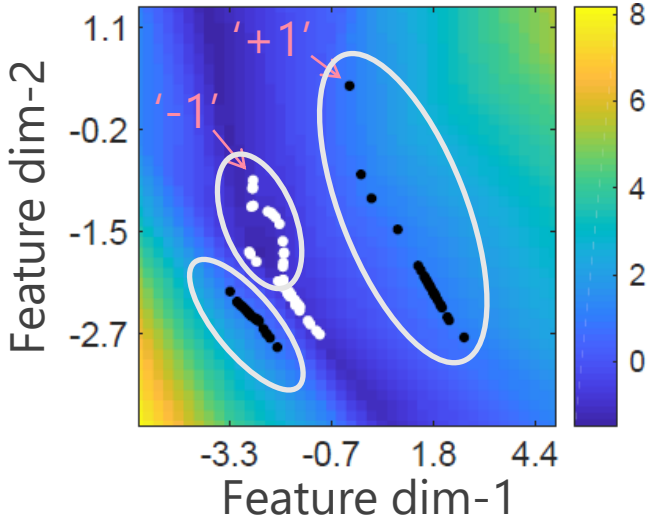
E2E optimization scheme flourishes. Is it always good?

E2E opt. $$(\phi^\star, \theta^\star) = \arg\min_{\phi,\theta} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t)\in\mathcal{D}} L\left(C_\theta\left(F_\phi(x)\right), t\right)$$



Feature extractor $F_{\phi^\star}$ adapts to a particular classifier C_θ.

Toy ex.)
2-class regression



color: C_θ value

Features may form
excessively complex distribution.

- Disjointed
- Split

FOCA: Feature-extractor Optimization through Classifier Anonymization

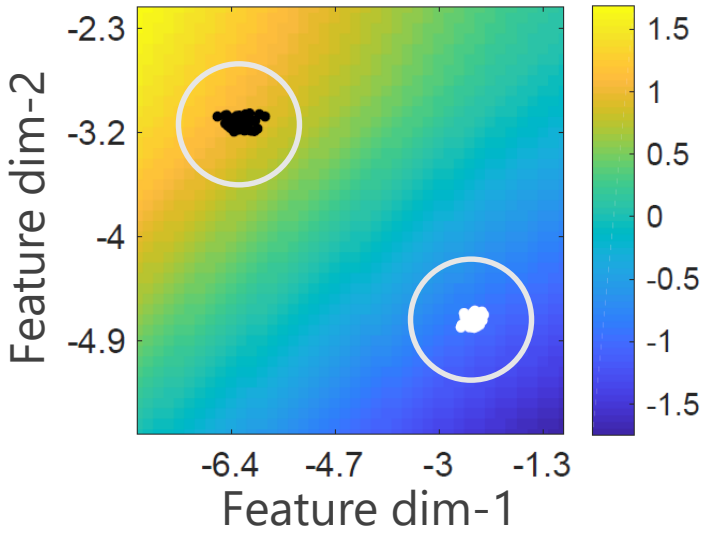
FOCA

$$\phi^\star = \arg\min_{\phi} \frac{1}{\|\mathcal{D}\|_0} \sum_{(x,t)\in\mathcal{D}} \mathbb{E}_{\theta\sim\Theta_\phi} L\left(C_\theta\left(F_\phi(x)\right), t\right)$$

Random weak classifier: $\theta\sim\Theta_\phi$

Want to know more about Θ_ϕ? Please come to the poster!

Feature extractor $F_{\phi^\star}$ adapts to a set of weak classifiers $\{C_\theta\}$.



Features form simple point-like distribution per class under some conditions.

Proposition about the point-like property

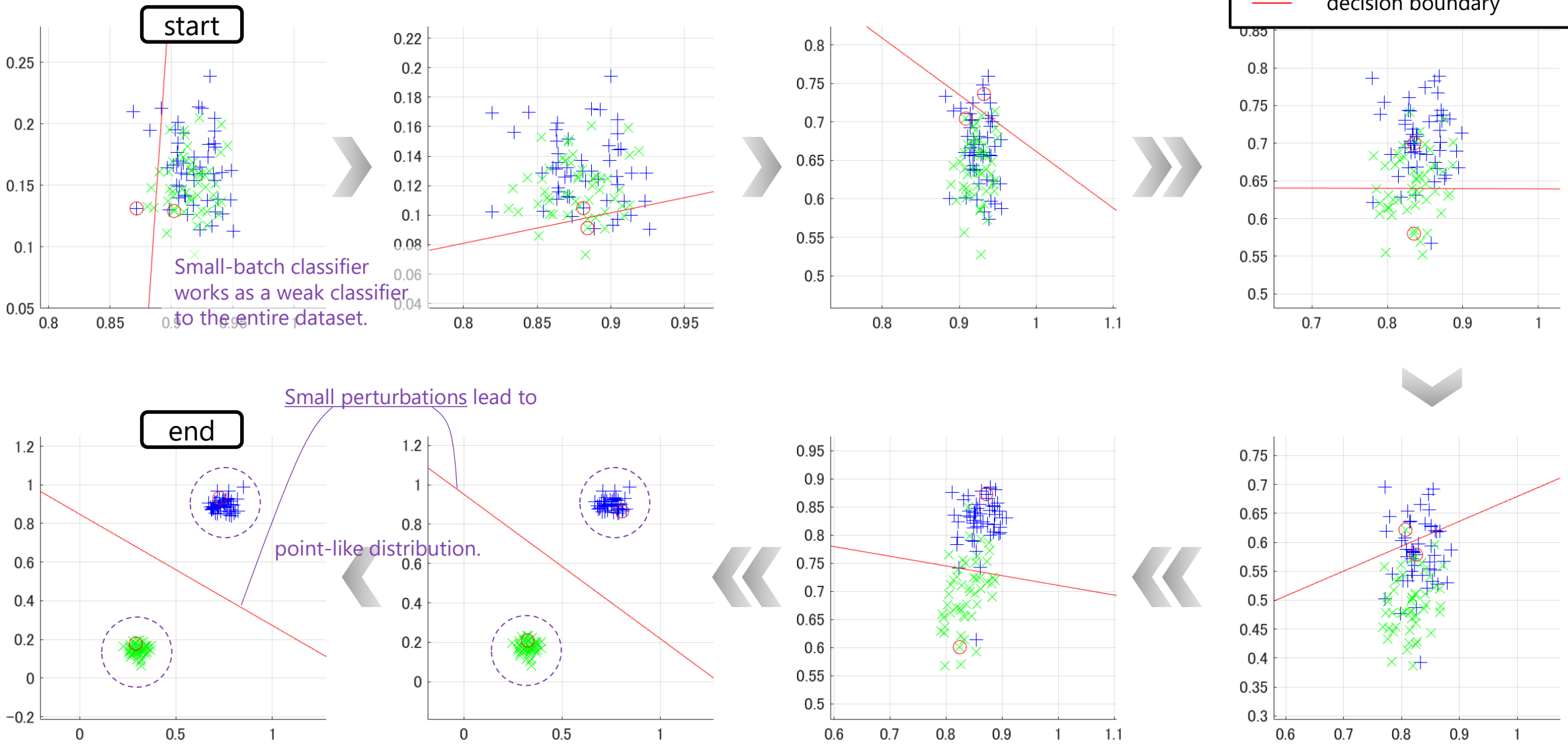
In words,

If feature extractor has an enough representation ability,
all input data of the same class are projected to a single point in the feature space in a class-separable way under certain conditions.

Please see the paper for the proof.

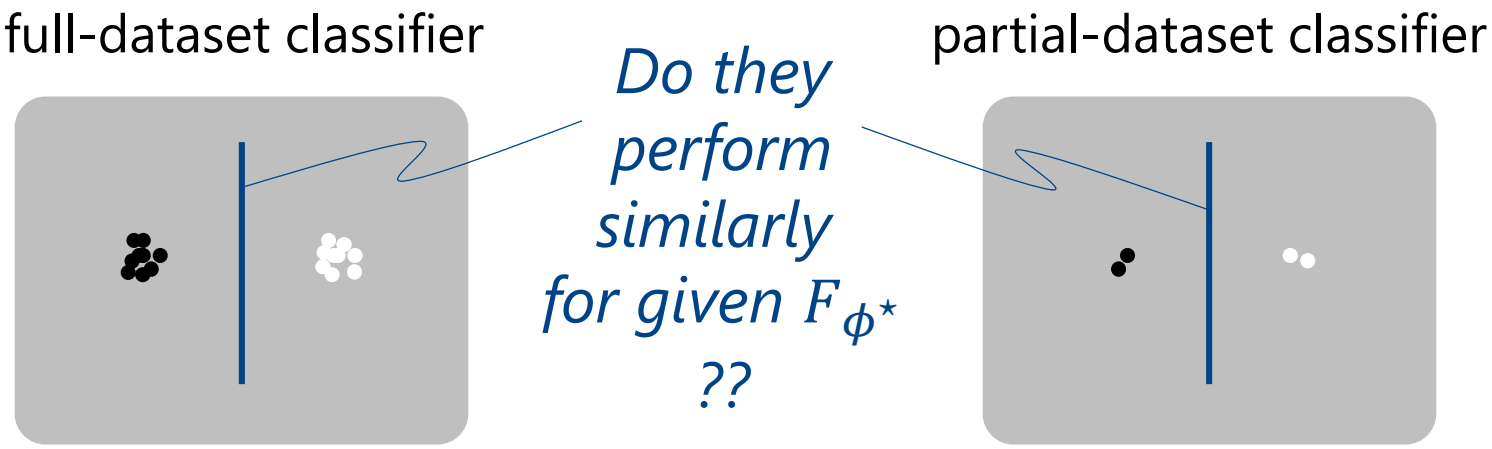
Toy problem demonstration

x-axis Feature dim. #1
y-axis Feature dim. #2
○ data used to generate classifier
— decision boundary

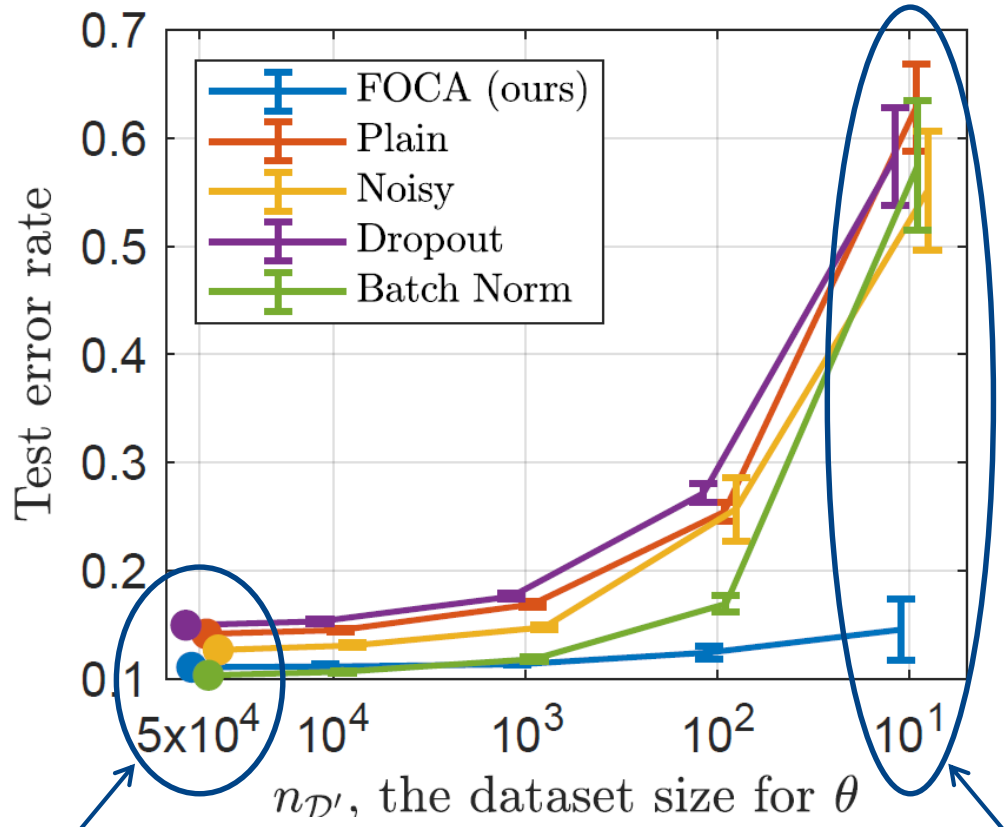


Experiment #1: partial-dataset training

Thing we wish to confirm:



Experiment #1: partial-dataset training



Performance gap

large for other methods

much smaller for FOCA

One indication of point-like property

classifier trained with large dataset

classifier trained with small dataset

(The same, fixed feature extractor is used within each method.)

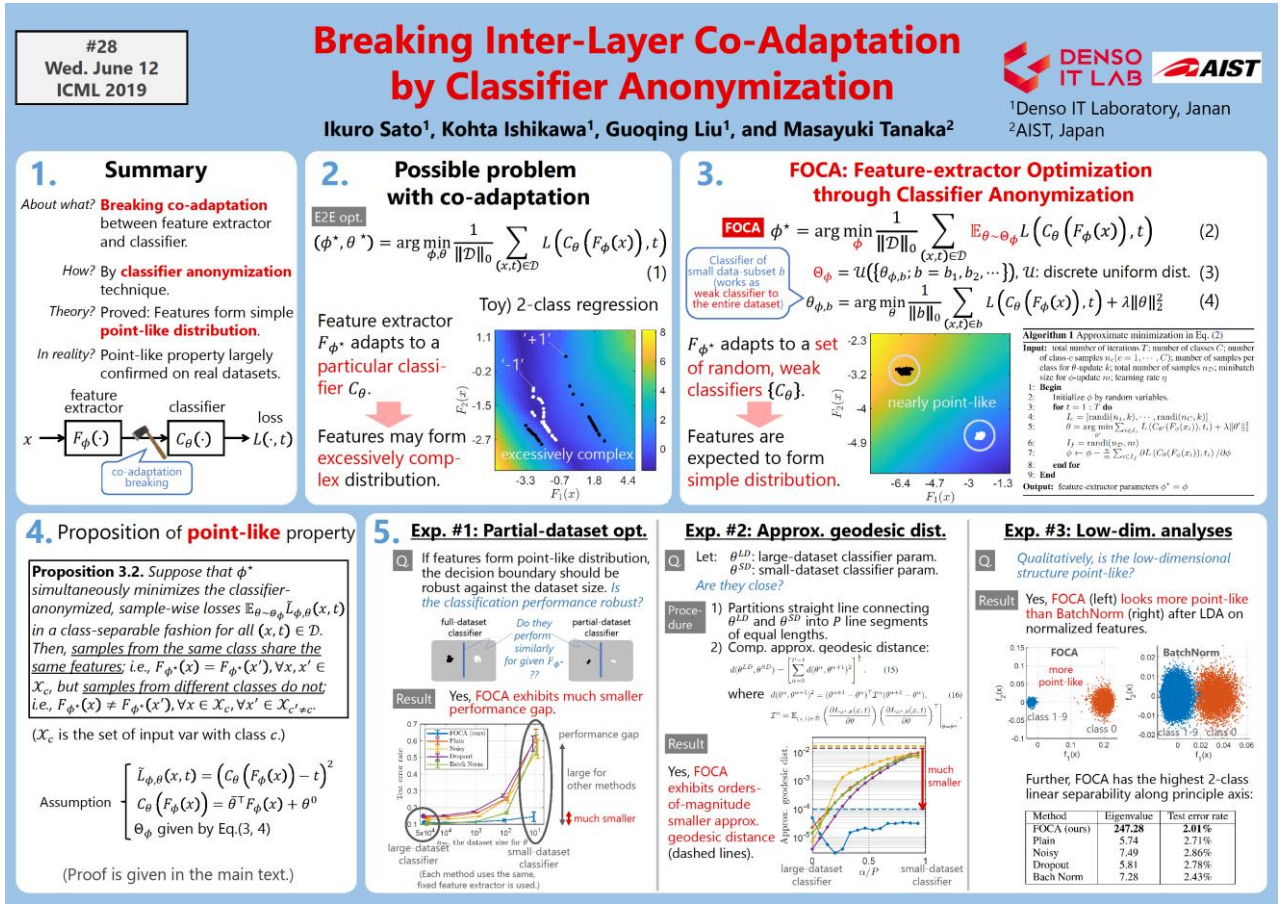
More experiments …

including:

- Approximate geodesic distance measurements between large- and small-dataset solutions
- Low-dimensional analyses

to further study the point-like property.

Poster #28 tonight



What? **Breaking co-adaptation** between feature extractor and classifier.

How? By **classifier anonymization**.

Theory? Proved: Features form simple **point-like distribution**.

Reality? Point-like property largely confirmed on real datasets.