UNIVERSITY OF
OXFORD

Ivans Lubenko and Andrew Ker

Steganalysis with Mismatched Covers:
Do Simple Classifiers Help?

6 September 2012
ACM Workshop on Multimedia and Security

PROBLEM

- Real world steganalysis is difficult:
  testing on images from unknown "source"
- *source = camera + pre-/postprocessing + ?*
  and influences image statistics
- Different sources form separate clusters in the feature space
- This results in reduced detection rates

POSSIBLE CAUSE

a) Our classifiers are undertrained: a limited training set does not allow for a model that generalises well to unknown data

b) Our models are too complex and overfit the image source

c) ???

(domain adaptation is hard)

OUR APPROACH:

- Train on more data

  up to 1,000,000 training examples

- Use CC-C300 features from [1]

  48,600 features, one of the best for JPEG domain

- Use simple (near-)linear classifiers

  1. Online Ensemble Average Perceptron
  2. Ensemble FLD

  both will be compared to Kernel SVM (**3.**)

[1] Jan Kodovsky, "Steganalysis in high dimensions: fusing classifiers built on random subspaces", 2011

AVERAGE PERCEPTRON

```
For each training example x:
    compute prediction:
```

$$y(x) = sign(w_{avg}^T x)$$ decision function

```
    if y(x) ≠ t, update weights w:
```

$$w_i = w_{i-1} + x_i t_i$$ true label of xi

```
    regularise via averaging:
```

average weight vector $$w_{avg} = w_{avg} + w_i$$

*This will actually be used in ensemble setting.

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

1. Matched training data

2. Mismatched training data

Evaluate on a sample steganalysis problem in JPEG domain:
cover vs nsF5 (0.05 bpnc)

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

1. Matched training data

26 sources x

6000 training / 2000 test images

KSVM	EFLD	OEAP
$\mu = 0.876$ $\sigma = 0.024$	$\mu = 0.892$ $\sigma = 0.026$	X*

*Requires min. 400,000 training examples to converge in the online setting.

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

2. Mismatched training data

   a) Less diverse training data

   b) More diverse training data

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

2. Mismatched training data

a) **Less** diverse training data:

KSVM: 6 random sources
EFLD: 20 random sources
OEAP: all 1,000 sources

x 1000 images

Fixed test data:

100 sources x 500 images

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

2. Mismatched training data

b) **More** diverse training data:

KSVM: 6,000 random images

EFLD: 20,000 random images

OEAP: all 1,000 sources x 1000 images

Fixed test data:

100 sources x 500 images

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

2. Mismatched training data

	KSVM (6000 samples)	EFLD (20 000 samples)	OEAP (1000 000 samples)
less diverse	$\mu = 0.804$ $\sigma = 0.071$	$\mu = 0.838$ $\sigma = 0.058$	
more diverse	$\mu = 0.809$ $\sigma = 0.039$	$\mu = 0.836$ $\sigma = 0.039$	$\mu = \mathbf{0.851}$ $\sigma = \mathbf{0.056}$

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

- Matched vs mismatched

	KSVM (6000 samples)	EFLD (20 000 samples)
matched	$\mu = 0.876$ $\sigma = 0.026$	$\mu = 0.892$ $\sigma = 0.024$
mismatched	$\mu = 0.809$ $\sigma = 0.039$	$\mu = 0.836$ $\sigma = 0.039$

performance drop

EXPERIMENTS:

Aim: to measure the performance drop between controlled data and real-world data

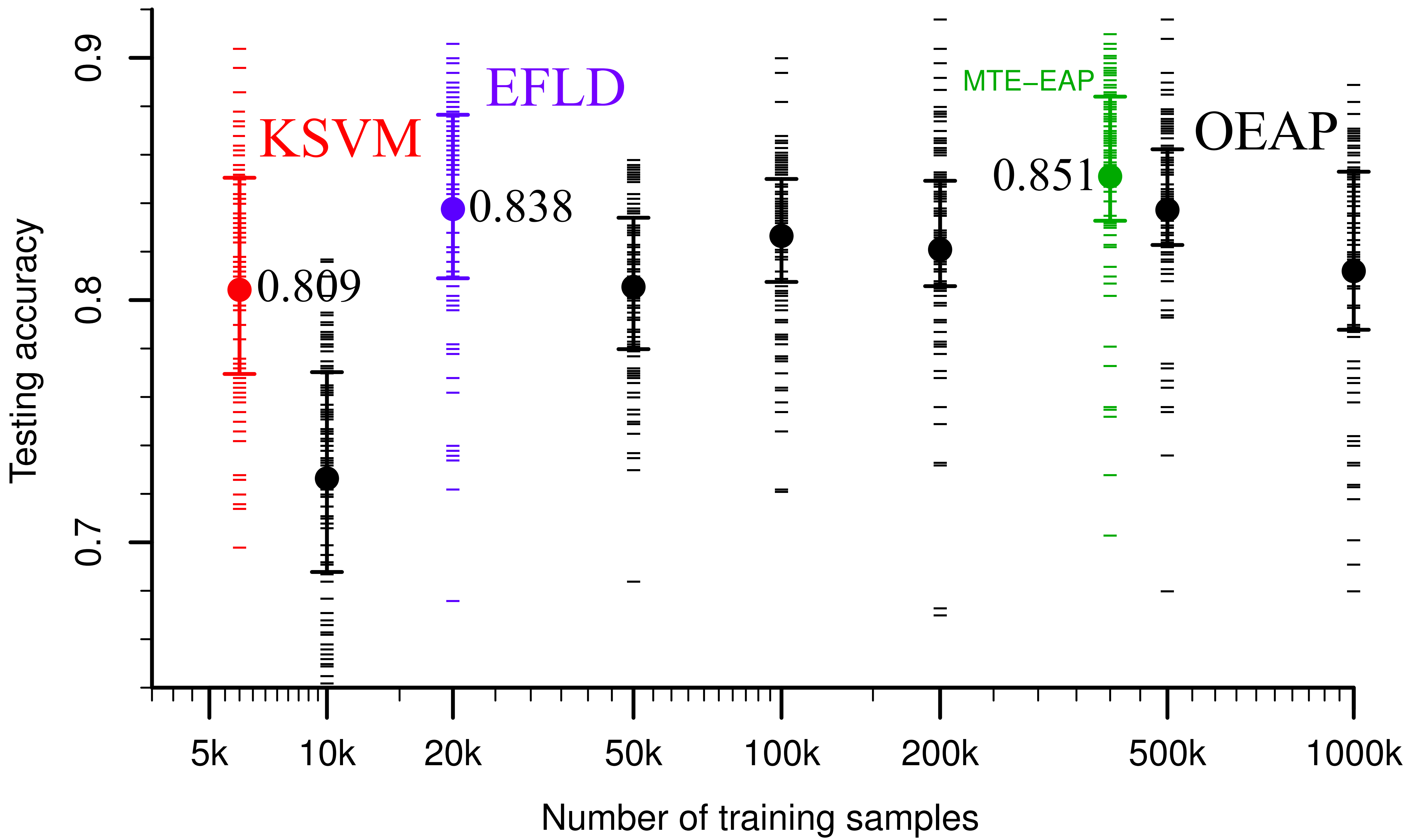
- Matched vs mismatched

	KSVM (6000 samples)	EFLD (20 000 samples)	OEAP (1000 000 samples)
matched	$\mu = 0.876$ $\sigma = 0.026$	$\mu = 0.892$ $\sigma = 0.024$	
mismatched	$\mu = 0.809$ $\sigma = 0.039$	$\mu = 0.836$ $\sigma = 0.039$	$\mu = 0.851$ $\sigma = 0.056$

statistically significant

performance drop

Mismatched data:



STILL POSSIBLE CAUSES:

a) Our classifiers are undertrained

training on more images allows for more variety of training data and improves accuracy but requires simpler classifiers

b) Our models are too complex and overfit the image source

using simpler models allows for more robust decision boundaries (e.g. EFLD in matched scenario) and hence also improves accuracy

- FUTURE DIRECTIONS:
  - Generalise the conclusions by studying more features/embedding schemes/?
  - Understand how much data is actually required for the classifier to converge.
  - Other non-linear online classifiers?