The effectiveness of subspace mapping techniques adapted to unlabeled samples from a global domain in mitigating sample selection bias

Timo van Hoorn
Email: T.F.R.vanHoorn@student.tudelft.nl
Supervisors: Joana de Pinho Gonçalves, Yasin Tepeli
EEMCS, Delft University of Technology, The Netherlands

1. Background

- **Subspace mapping techniques** aim to find a common subspace between the source and the target domain.
- Examples of subspace mapping techniques are **subspace alignment** (SA) [1] and **transfer component analysis** (TCA) [2].

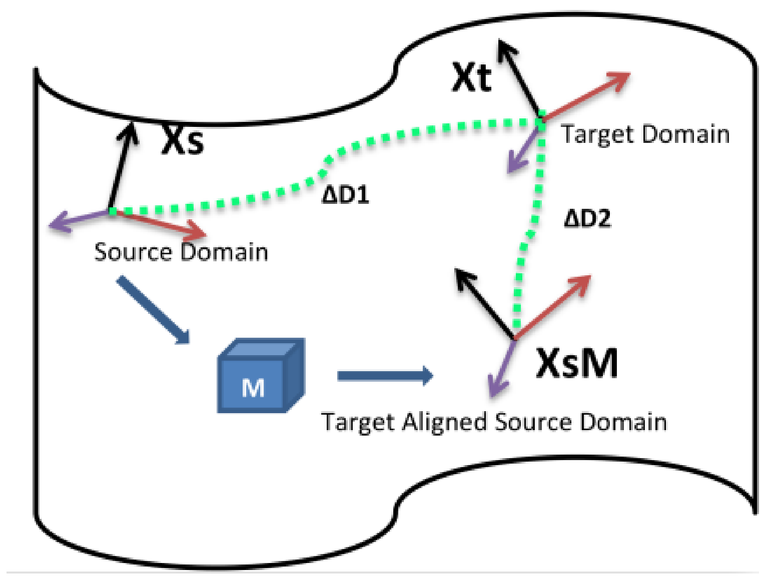

Figure 1. Visualization of Subspace Alignment that aligns the source domain with the target domain [1].

2. Research question

Main research question:

- How effective are subspace mapping techniques in mitigating sample selection bias?
- **Assumption**: Only unlabeled samples from an underlying global domain and **no** samples from the target domain

References

[1] Basura Fernando, Amaury Habrard, Marc Sebban and Tinne Tuytelaars. Subspace Alignment For Domain Adaptation. 2014
[2] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok and Qiang Yang. Domain Adaptation via Transfer Component Analysis.

3. Methodology

The method for getting results consists of three parts:

1. **Generate** training, testing, and global datasets.
2. **Train** SA, and TCA models using the training and global datasets.
3. **Evaluate** the models using the testing set.

A biasing technique is used to bias the **training** data:

- When randomly selecting samples from the original dataset, we put more weight on samples that are close to a random **biasing data point** in the space.
- A **biasing factor** is used to determine the degree of bias.
- The **higher** the biasing factor, the **more** weight is put on samples close to the biasing data point.

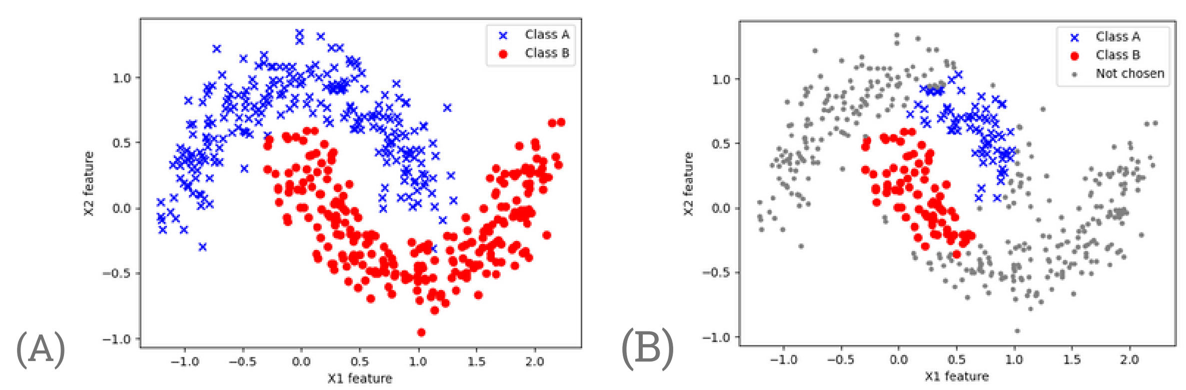

Figure 2. (A) A plot containing all samples in a synthetic data set. (B) A plot containing a biased data set drawn from the data set in A using the biasing technique described above. The biasing data point used in the biasing technique is (0.5, 0.5).

4. Experimental setup

- Experiments on **four** factors: training sample size, feature size, proxy A-distance, and initial data set.
- Test SA and TCA using **k-nearest neighbors** (KNN) and **logistic regression** (LR) as **estimator** parameters.
- Measure the lowest, highest, and mean accuracy on 10 different train, test, global splits, and biasing data points.

5. Results

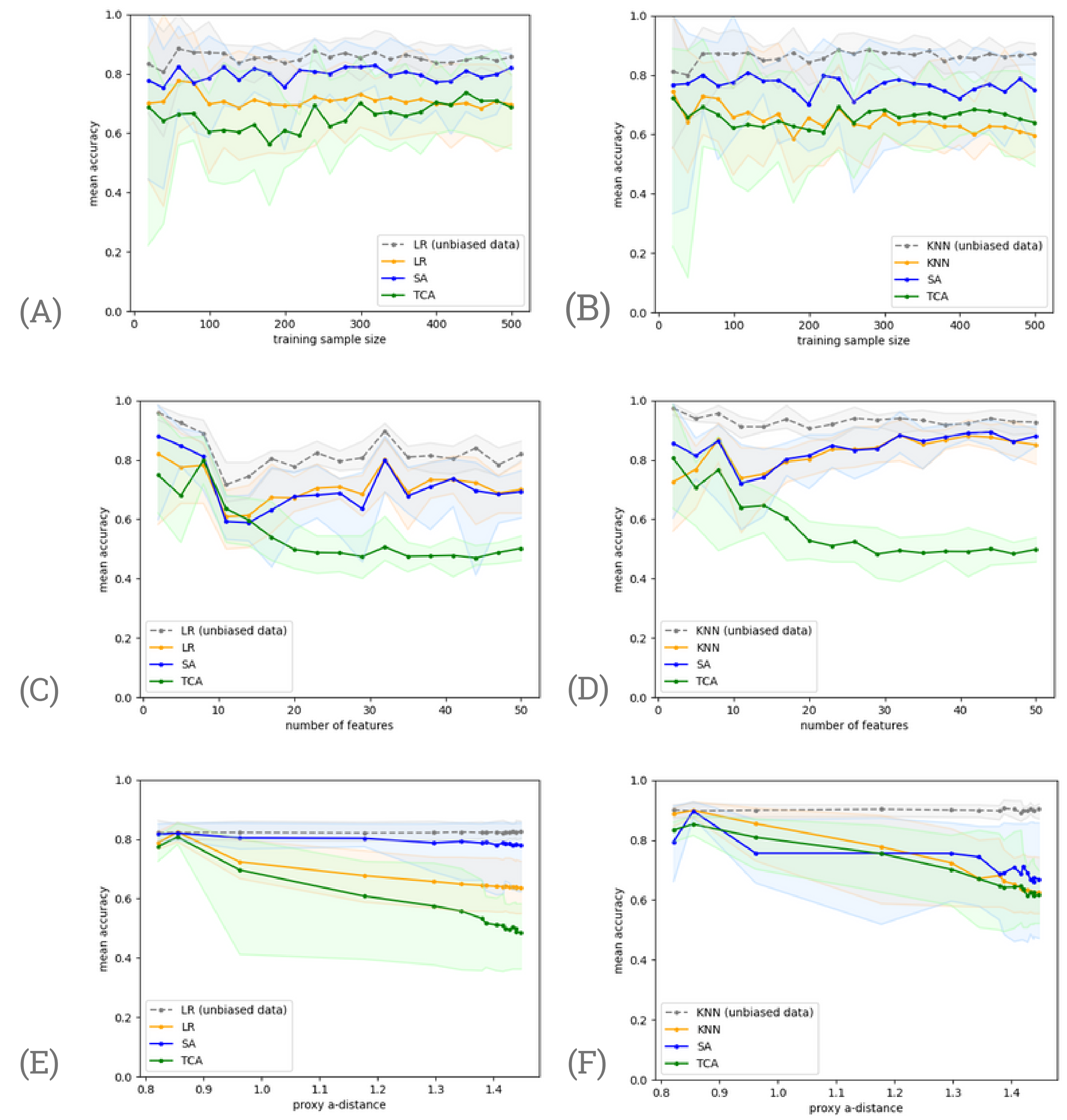

Figure 3. Mean, lowest and highest accuracies. The gray dotted line indicates a classifier trained on an unbiased training set. (A) (B) Varying training sample sizes. (C) (D) Varying numbers of features. (E) (F) Varying proxy A-distances. (A) (C) (E) SA and TCA with estimator set to LR. (B) (D) (F) SA and TCA with estimator set to KNN.

6. Conclusions

- SA effectively mitigates sample selection bias on data sets with a **low** number of features and with a **high** distance between the source and target domain.
- TCA is more effective with **more** training samples and on data sets with only a **few** features where the distance between the source and target domain is **not** too big.