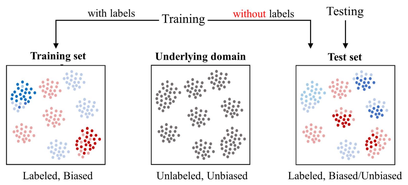
Evaluating the Effectiveness of Importance Weighting Techniques in Mitigating Sample Selection Bias

Andrei Camil Tociu (A.C.Tociu@student.tudelft.nl)
EEMCS faculty, Delft University of Technology, The Netherlands

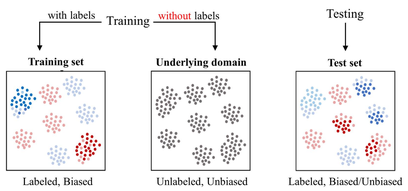
1. Introduction

Supervised machine learning classifiers often assume that data in the train and test sets follow the same probability distribution. ***Sample selection bias*** occurs in many real-life situations and causes this assumption to fail, which diminishes the prediction ability of the classifier.
Importance weighting corrects the discrepancy in the two distributions by assigning weights to the cost of error of each train point. It does so by using unlabeled data from the dataset to which it adapts.

Literature approach: adapt train set to a particular test set.
→ non-generalizable to different test sets, fails for small test set sizes
New approach: adapt train set to the underlying domain of the data.



a) from the specific test set distribution



b) from the underlying domain distribution

Figure 1: Learning approaches for domain adaptation based on the provenance of the unlabeled train data. Our research uses approach (b).

Research Question: ***How effective is importance weighting in mitigating sample selection bias when the unlabeled data is sourced from the underlying domain instead of a particular test set ?***

Scenarios commonly known for posing difficulties to importance weighting:
- Unequal conditional probabilities of source and target domains
  i.e. $P_S(y|x) \neq P_T(y|x)$
- Small train sample size
- High-dimensional data

4. Conclusion

1) Importance weighting is **not a one-size-fits-all solution**.

2) Theoretical **performance bounds** for importance weighting **do not hold** anymore when the **fundamental assumption** of equal conditional probabilities is **violated**.

3) a) Some methods can still significantly **improve classification** even when the **fundamental assumption** made by importance weighting is **violated**.

3) b) Degree of overlap between classes seems to influence adaptation ability in the scenario above.

4) A **small** train set **size negatively influences** adaptation ability. However, the "intensity" of the **sampling bias** is **equally important**.

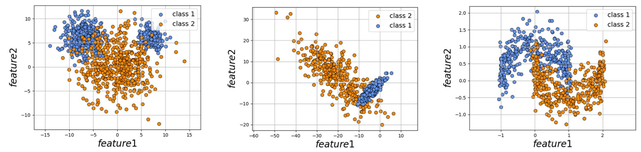
5) **Curse of dimensionality is not universal** for importance weighting.

2. Methodology

Unequal Conditional Probabilities

Datasets: synthetic, binary classification, balanced, two features

Sampling bias depends on label $\boldsymbol{y}$. Vary class proportions in the train set from 50-50% to 2-98% .



(a) Three unbalanced clusters (b) Overlapping linear clusters (c) Rotated moons

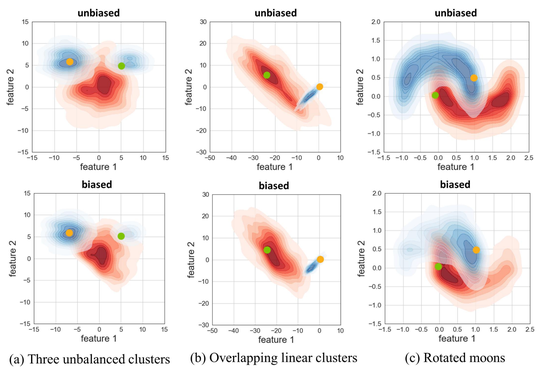
Figure 2: A visualization of the synthetic classification datasets used in the experiments. Class 1 is marked in blue and class 2 in red.

Small Train Sample Size

Sampling bias depends on feature vector $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$. Select samples closer to certain points with coordinates $(\Delta_{x_1}, \Delta_{x_2})$ in the 2D space described by x:

$$P(s = 1|x) = e^{-b * (|x_1 - \Delta_{x_1}| + |x_2 - \Delta_{x_2}|)}$$

Reduce train set size from 100% to 2%.



(a) Three unbalanced clusters (b) Overlapping linear clusters (c) Rotated moons

Figure 3: Probability density function of the synthetic datasets, before and after applying the biasing scheme. The points with coordinates $(\Delta_{x_1}, \Delta_{x_2})$ used in the biasing scheme are marked in yellow for class 1 (blue) and in green for class 2 (red).

High-Dimensional Data

Generate random binary classification, balanced **datasets** for different feature sizes (10, 20, 30, 40, 50).
Use multiple sample(s)-feature(f) functions to ensure an adequate train set size:
$s = f * 50$, $s = f * 100$, $s = f^2 * 5$.

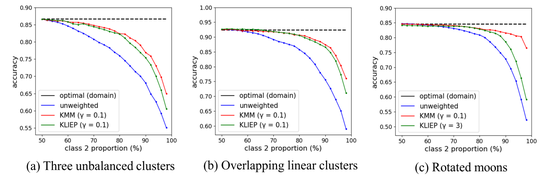
Sampling bias based on:
a) the most important feature only
b) all features while maintaining class balance
c) all features without maintained class balance

Metric: *percentual domain adaptation (%)* = $100 * \frac{Acc_{IW} - Acc_{unweighted}}{Acc_{optimal} - Acc_{unweighted}}$

3. Results

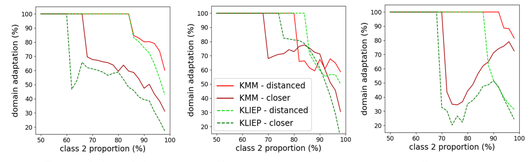
Evaluation framework applied to two importance weighting techniques:
Kernel Mean Matching (KMM) and Kullback-Leibler Estimation Procedure (KLIEP).

- KMM and KLIEP initially perform on par with the optimum and then significantly outperform the unweighted classifier on all datasets.
- Their performance seems to be affected by the overlapping class regions (→ majority class outweighs the minority class).
  - Adaptation ability for more distanced clusters is generally better.
  - Adaptation ability for different train set sizes does not change.
- Assigning class weights improves accuracy of both the weighted and unweighted classifiers up to the optimum.



(a) Three unbalanced clusters (b) Overlapping linear clusters (c) Rotated moons

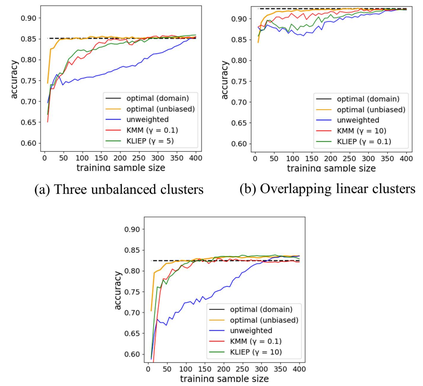
Figure 4: Classification performance on a label-based sampling scheme.



(a) Three unbalanced clusters (b) Overlapping linear clusters (c) Rotated moons

Figure 5: Domain adaptation performance for more versus less overlapping classes.

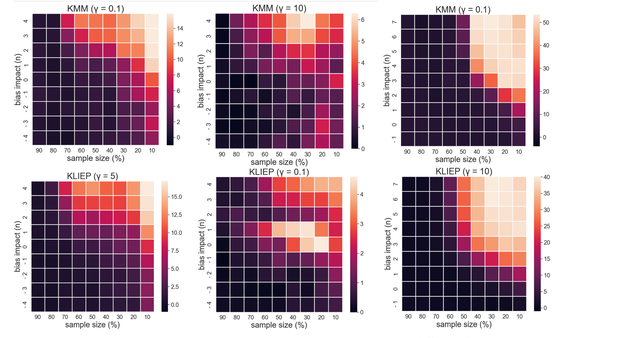
- KMM and KLIEP initially perform on par with the optimum and then significantly outperform the unweighted classifier on all datasets.
- The performance of KMM and KLIEP greatly degrades with a decreasing train sample size.
  The effect of sample size on the accuracy curve is non-linear: first close to optimum, then slow decline, and finally sharp decrease.
- Sampling bias intensity is an important confounding factor of the decrease in performance under a reduced train sample size.



(a) Three unbalanced clusters (b) Overlapping linear clusters

(c) Rotated moons

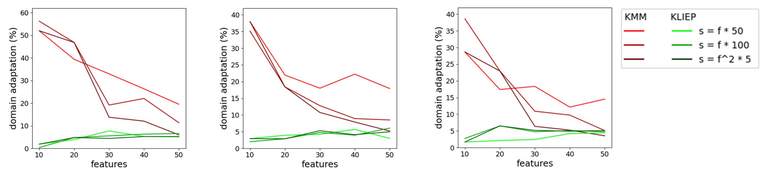
Figure 6: Classification performance on a feature-based sampling scheme for varying train sample sizes.



(a) Three unbalanced clusters (b) Overlapping linear clusters (c) Rotated moons

Figure 7: Accuracy difference between the optimal (domain) and importance-weighted classifiers as a function of the bias impact factor n (used for varying the values of b) and the proportion of the original train set size (400 samples). The original values of b used in the sampling scheme are varied as follows: (a) $b * 1.5^n$ for class 1 and $b * 1.3^n$ for class 2; (b) $b * 2.3^n$ for both classes; (c) $b * 2^n$ for both classes. Lighter colors indicate higher differences, so poorer adaptation performance.

- KMM greatly affected by high-dimensional data for all three types of bias.
  - "Strength" of the biased sampling scheme is not necessarily correlated with the impact of high-dimensionality.
  - A biased sampling scheme including labels does not impact much more than one based solely on the feature space.
- KLIEP does not show performance fluctuations under high-dimensional data on any of the biasing schemes.
  However, its performance is generally poor throughout the experiment (scores in range [0.39, 7.74]).



(a) most relevant feature (b) all features (balanced classes) (c) all features (imbalanced classes)

Figure 8: Domain adaptation performance for different feature dimensions. Three sampling schemes are used.