Domain Adaptation and Covariate Shift – A Literature Review

Speaker: Maximilian Autenrieth[1],
Supervisor: David A. van Dyk[1]
Co-Supervisor: David Stenning[2]

Imperial College London[1], Simon Fraser University[2]

February 25, 2020

Causes and cases of covariate shift and domain adaptation

- **Astronomy:** Spectroscopical follow-up of astronomical sources not at random. Most promising objects are selected.
- **Medical Imaging:** Radiologists manually annotate pathologies (e.g. in MRI's). Mechanical configurations vary between medical centers.
- **Natural language processing:** Annotated training data (e.g. Wall street journal) is highly specialized.
- **Robotics:** Supplementary simulated training data is added to support real-life predictions (e.g. pedestrian detection).
- **Fairness aware machine learning:** Ensure that automated decision-making systems do not discriminate people based on certain attributes (e.g. gender, race).
- **Knowledge transfer:** Improve speech recognition based on natural language processing data.

Causes and cases of covariate shift and domain adaptation:

- **Computer Vision:** Use web-crawler collected product images to classify real-world collected images.

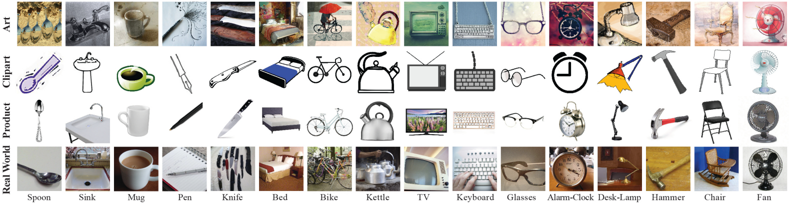
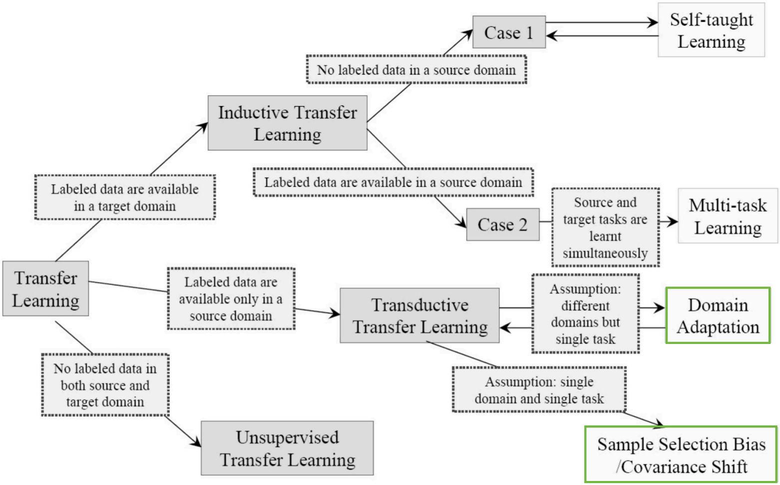


Figure 0.1: Sample-images from the Office-Home dataset Venkateswara et al. (2017), consisting of images from four domains: Art, Clipart, Product and Real-World.

1. Categorization and Terminology of Domain Adaptation and Covariate Shift:
2. Covariate Shift and Sample Selection Bias:
3. Unsupervised and Semi-supervised Domain Adaptation:
4. Propensity Score Methodology:
5. Covariate Shift in Astronomy – Improving Supernova Type Ia Classification:

Categorization and Terminology:



Source: Pan and Yang (2009)

Definitions and Notation:

Let $\mathcal{X} \subset \mathbb{R}^F$, $F > 0$, be the feature space and $\mathcal{Y}$ the label space with $K > 1$ classes, or a subset of $\mathbb{R}$ in the regression case. Different domains are defined as different probability distributions $p(x, y)$ over the same feature-label space pair $\mathcal{X} \times \mathcal{Y}$ (Kouw and Loog 2019).

Unsupervised Domain Adaptation:

- Source data: $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labelled samples, from joint distribution p_S (Domain $\mathcal{D}_S$),
- Target data: $D_T = \{x_j^t\}_{i=1}^{n_t}$ with n_t unlabelled samples, from joint distribution p_T (Domain $\mathcal{D}_T$),

with $p_S(x, y) \neq p_T(x, y)$.

Semi-Supervised Domain Adaptation:

- At least one target label y^t is given.

Definitions:

Definition 1.1

Covariate shift appears only in $\mathcal{X} \to \mathcal{Y}$ problems, and is defined as case where $p_S(y|x) = p_T(y|x)$ and $p_S(x) \neq p_T(x)$.

Definition 1.2

Prior (target) shift appears only in $\mathcal{Y} \to \mathcal{X}$ problems, and is defined as case where $p_S(x|y) = p_T(x|y)$ and $p_S(y) \neq p_T(y)$.

Definition 1.3

Concept shift is defined as

- $p_S(y|x) \neq p_T(y|x)$ and $p_S(x) = p_T(x)$ in $\mathcal{X} \to \mathcal{Y}$ problems
- $p_S(x|y) \neq p_T(x|y)$ and $p_S(y) = p_T(y)$ in $\mathcal{Y} \to \mathcal{X}$ problems

Definitions from Moreno-Torres et al. (2012). Prior and Concept shift is e.g. discussed in Widmer and Kubat (1996); Zhang et al. (2013).

Definitions:

Let $f : \mathcal{X} \to \mathbb{R}^K$ our training function, and f an element of the hypothesis space $\mathcal{H}$. Then,

- $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ is the loss function
- $\mathcal{R}(f) = \mathbb{E}[\ell(f(x), y)]$ is the risk function

Aim: Minimize the risk $\mathcal{R}_T$ on the target domain $\mathcal{D}_t$, given labelled source data D_S and unlabelled target data D_T.

Sample/loss Reweighting

Covariate shift: $p_S(y|x) = p_T(y|x)$ and $p_S(x) \neq p_T(x)$

Proposition 1 (Bickel et al. (2009); Shimodaira (2000))

If the support of $p_T(x)$ is contained in $p_S(x)$, the expected loss w.r.t. $\mathcal{D}_T$ equals the expected loss w.r.t. $\mathcal{D}_S$ with weights $w(x) = p_T(x)/p_S(x)$ for the loss incurred by each x,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_T}\left[\ell(f(x),y)\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}_S}\left[\frac{p_T(x)}{p_S(x)}\ell(f(x),y)\right]$$

This follows from:

$$\begin{aligned}\mathcal{R}_T(f) &= \sum_{y\in Y}\int_{\mathcal{X}} \ell(f(x),y)\frac{p_T(x,y)}{p_S(x,y)}p_S(x,y)dx \\ &= \sum_{y\in Y}\int_{\mathcal{X}} \ell(f(x),y)\frac{p_T(y|x)}{p_S(y|x)}\frac{p_T(x)}{p_S(x)}p_S(x,y)dx\end{aligned}$$

Maximum weighted log likelihood (Shimodaira 2000):

Assumptions:

1. Covariate shift: $p_S(x) \neq p_T(x)$
2. Model misspecification

For sufficiently large n, Shimodaira (2000) proposes weighted maximum likelihood estimation:

$$L_w(\theta|x, y) := \sum_{t=1}^{n} w(x) \log p(y|x, \theta), \quad \theta \in \Theta.$$

For moderate sample sizes:

$$w_\alpha = \left(\frac{p_T(x)}{p_S(x)}\right)^\alpha, \quad \alpha \in [0, 1].$$

Optimal α can be determined by a variant of Akaike's information criterion (Akaike 1974).

Example of Maximum Weighted Log-Likelihood:

- $X_S \sim N(0.5, 0.5^2)$ and $X_T \sim N(0, 0.3^2)$
- $y = -x + x^3 + \epsilon$, with $\epsilon \sim N(0, 0.3^2)$
- $w(x) = p_T(x)/p_S(x) \propto \exp\left(-(x - \bar{\mu})^2/(2\bar{\tau})\right)$

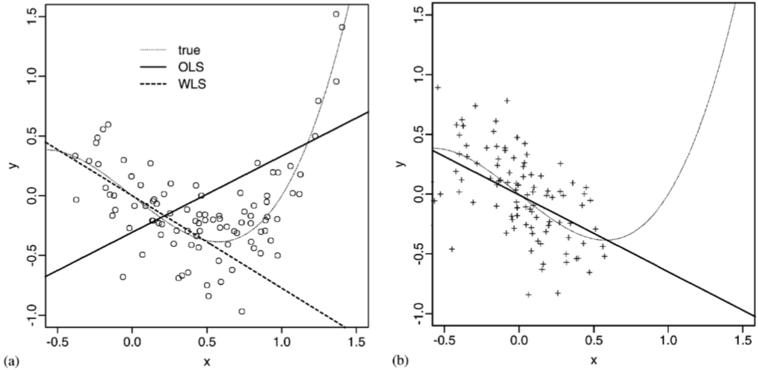


Figure 2.1: (Shimodaira 2000) Polynomial regression fitting with degree $d = 1$. a) n=100 samples from $p_S(X_S)$. b) n=100 from $p_T(X_T)$.

Importance Weighted Cross-Validation (Sugiyama et al. 2007):

Divide training data $D_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ into k disjoint, equally-sized subsets $\{D_S^i\}_{i=1}^{k}$. Let $f_{D_s^j}(x)$ be a function learned from $\{D_S^i\}_{i \neq j}$, the weighted *k-fold* cross-validation estimate of the risk $\mathcal{R}^{(n)}(f)$ is given by:

$$\hat{\mathcal{R}}_{WCV}^{(n)} := \frac{1}{k} \sum_{j=1}^{n} \frac{1}{|D_s^j|} \sum_{(x,y) \in D_s^j} \frac{p_T(x)}{p_S(x)} \ell(f_{D_s^j}(x), y)$$

For weighted leave-one-out-CV (LOOWCV) it holds that (Sugiyama et al. 2007):

$$\mathbb{E}_{(x,y)} \left[\hat{\mathcal{R}}_{LOOWCV}^{(n)} \right] = \mathcal{R}^{(n_t - 1)},$$

where $\mathcal{R}^{(n_t-1)}$ is the risk on $(n_t - 1)$ target samples D_T.

Covariate Shift with Sample Selection Bias:

- Sample selection bias is a widely studied issue (Heckman 1977; Little and Rubin 2019; Rosenbaum and Rubin 1983).
- Zadrozny (2004) introduce sample selection bias in a general machine learning framework:

Bias scenario:

- Examples (x, y, s) are drawn from a domain $\mathcal{D}$, with feature-label-selection space $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$.
- $S \in \mathcal{S}$ is a latent, binary indicator variable that controls the training set selection ($s = 1$).
- S depends on X, but S is independent of Y given X ($P(s = 1|x, y) = P(s = 1|x)$).

Bias Correction (Zadrozny 2004)

Proposition 2 (Bias Correction (Zadrozny 2004))

For any distribution $\mathcal{D}$, for all classifiers f, for any loss function $\ell = \ell(f(x), y)$, if we assume that $P(s = 1|x, y) = P(s = 1|x)$ (that is, s and y are independent given x), then

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(f(x), y)\right] = \mathbb{E}_{(x,y)\sim\hat{\mathcal{D}}}\left[\ell(f(x), y)|s = 1\right],$$

$$\text{with } \hat{\mathcal{D}}(x, y, s) := \frac{P(s = 1)}{P(s = 1|x)}\mathcal{D}(x, y, s)$$

- We can minimize the expected target loss, by drawing samples from $\hat{\mathcal{D}}$.

Local and Global Learners (Zadrozny 2004):

- **Local:** The output of the learner depends asymptotically only on $P(y|x)$.
- **Global:** The output of the learner depends asymptotically both on $P(x)$ and on $P(y|x)$.

Local	Bayesian Classifier Logistic Regression (correcly specified) Hard margin SVM
Global	Naive Bayes Decision Trees Soft margin SVM

Importance Estimation:

Proposed methods to estimate $w(x) = \left(\frac{p_T(x)}{p_S(x)}\right)$

- Kernel density estimation (Shimodaira 2000)
- Kernel Mean Matching – in reproducing kernel Hilbert space (Huang et al. 2007)
- Logistic regression (Bickel and Scheffer 2007; Zadrozny 2004)
- Kernel Logistic Regression – joint optimization problem (Bickel et al. 2009)
- Kullback-Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al. 2008)
- KLIEP extensions (Tsuboi et al. 2009) and unconstrained least-squares importance fitting (uLSIF) (Kanamori et al. 2009; Umer et al. 2019)

Domain Dissimilarity and Generalization Error:

Domain dissimilarity is measured to estimate generalization error across domains. Rényi divergence (Van Erven and Harremos 2014):

$$\mathcal{Q}_{\mathcal{R}^\alpha}[p_T, p_s] = \frac{1}{\alpha - 1} \log_2 \int_{\mathcal{X}} \frac{p_T^\alpha(x)}{p_S^{\alpha-1}(x)} dx$$

Other metrics: Kullback-Leibler divergence, Wasserstein metric, Kolmogorov-Smirnoff statistic (Cover and Thomas 2012; Mahmud 2009)

Generalization error: With probability $1 - \delta$ for $\delta > 0$ (Cortes et al. 2010)

$$|e_T(f) - \hat{e}_W(f)| \leq \; 2^{5/4} \; 2^{\mathcal{Q}_{\mathcal{R}^2}[p_T, p_s]/2} \; \sqrt[3/8]{\frac{c}{n} \log \frac{2ne}{c} + \frac{1}{n} \log \frac{4}{\delta}},$$

with empirically weighted source error $\hat{e}_W(f)$, corresponding to a 0/1-loss, and c, the *pseudo-dimension* of the hypothesis space (Kouw and Loog 2019; Vidyasagar 2002).

Semi-supervised Domain Adaptation:

Given labelled source data D_s and target data $D_T = \{x_j^t\}_{i=1}^{n_t}$, with n_l labelled examples and $n_t - n_l$ unlabelled examples, $n_l << n_t$. Feature space is $\mathcal{X} = \mathbb{R}^F$.

Daumé III (2009): "Frustratingly easy Domain Adaptation"

- Augment the input space by $\mathcal{X}_a = \mathbb{R}^{3F}$ and define mappings $\Phi^s, \Phi^t : \mathcal{X} \rightarrow \mathcal{X}_a$ given by:

$$\Phi^s(x) = \langle x, x, \mathbf{0} \rangle, \quad \Phi^t(x) = \langle x, \mathbf{0}, x \rangle$$

  $\mathbf{0} = \langle 0, 0, \ldots, 0 \rangle \in \mathbb{R}^F$ is the zero vector.
- Train a multilayer-perceptron on the augmented data and predict on the unlabelled target samples.

Unsupervised Domain Adaptation:

Domain adaptation with deep neural networks (Long et al. 2015):

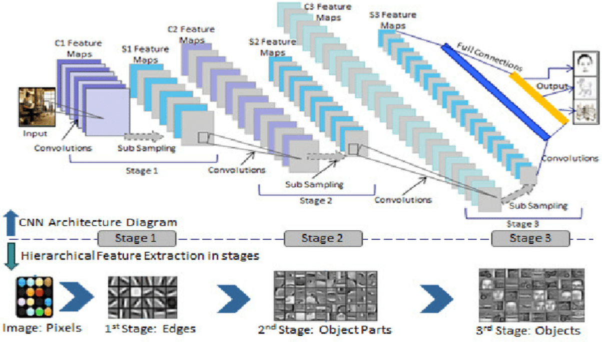


Figure 3.1: Architecture and learning stages of a deep convolutional neural network. (Katole et al. 2015)

- In the first layers DNNs learn general features, not specific to a particular task (Yosinski et al. 2014).

Domain Adaptation with Deep Neural Networks:

Idea: Jointly train the DNN on labelled source data and match moments of deep source and target feature maps:

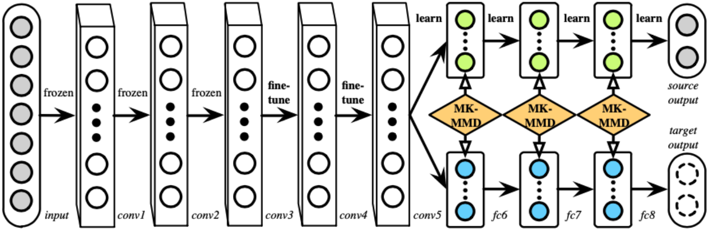


Figure 3.2: Deep Adaptation Network (DAN) (Long et al. 2015)

DAN risk function, with $\lambda > 0$, $l_1 = 6$ and $l_2 = 8$:

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \sum_{m=l_1}^{l_2} d_k^2(\mathcal{D}_s^m, \mathcal{D}_t^m)$$

Deep Adaptation Network (DAN):

Multiple kernel maximum mean discrepancies (Gretton et al. 2012):

$$d_k^2(p, q) := \| \mathbb{E}_p[\phi(x^s)] - \mathbb{E}_p[\phi(x^t)] \|^2_{\mathcal{H}_k},$$

with $p = q$ *iff* $d_k^2(p, q) = 0$, and ϕ denotes the deep feature map.

Proposition 3 (Long et al. (2015))

Let $f \in \mathcal{H}$ *be a hypothesis,* $e_S(f)$ *and* $e_T(f)$ *be the expected risks of source and target respectively, then*

$$|e_T(f) - \hat{e}_S(f)| \leq 2d_k(p, q) + C,$$

where C *is a constant for the complexity of hypothesis space and the risk of an ideal hypothesis for both domains.*

Further influential approaches:

(Daume III and Marcu 2006):

- Break source and target domain into three underlying distributions: q_S, q_T and q_G.
- Employ conditional expectation maximization to split into source specific, target specific and general information.
- Use general and target samples for prediction on unlabelled set.

Tzeng et al. (2017)

- Adversarial Discriminative Domain Adaptation

Ganin and Lempitsky (2014)

- Unsupervised Domain Adaptation by Backpropagation

Propensity Score Methods in Observational Studies:

- Rosenbaum and Rubin (1983) introduce propensity score:

$$e(X) = P(Z = 1|X).$$

- Treatment assignment Z is strongly ignorable, if

$$\text{(i)} \;\; (Y_1, Y_0) \perp\!\!\!\perp Z|X \qquad \text{and} \qquad \text{(ii)} \;\; 0 < e(X) < 1. \qquad (1)$$

- If (1) holds, PS is a balancing score
  ⇒ conditional on the PS, treatment effect estimates unbiased
- Four PS methods:
  Inverse probability of treatment weighting (IPTW),
  PS covariate adjustment, stratification and matching on PS

Inverse Probability of Treatment Weighting (IPTW)

- Weights in IPTW:

$$w_{ATE} = \frac{Z}{e(X)} + \frac{1-Z}{1-e(X)} \qquad \text{and} \qquad w_{ATT} = Z + \frac{e(X)(1-Z)}{1-e(X)}.$$

- Lunceford and Davidian (2004) introduce a consistent average treatment effect estimator

$$\hat{\Delta}_{IPTW2} = \Big(\sum_{i=1}^{n} \frac{Z_i}{e_i(X)}\Big)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{e_i(X)} - \Big(\sum_{i=1}^{n} \frac{1-Z_i}{1-e_i(X)}\Big)^{-1} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-e_i(X)}.$$

1. Categorization and Terminology of Domain Adaptation and Covariate Shift:
2. Covariate Shift and Sample Selection Bias:
3. Unsupervised and Semi-supervised Domain Adaptation:
4. Propensity Score Methodology:
5. Covariate Shift in Astronomy – Improving Supernova Type Ia Classification:

STACCATO – Supernova Photometric Classification with Biased Training sets

Data: "Supernova photometric classification challenge" (Kessler et al. 2010b)

- 17,330 simulated supernovae of type Ia, Ib, Ic and II.
- For each SN, light curve observations are given in four color bands C= (g,r,i,z).
- Training set: 1,217 spectroscopically confirmed SNe with known types
- Test set: 16,113 SNe with unknown types and photometric information alone

Approach: STACCATO - 'Synthetically Augmented Light curve Classification' Revsbech et al. (2018)

- Interpolation of light curves with with Gaussian processes
- Compute diffusion map for feature extraction (Richards et al. 2012), then classify the samples with a random forest.

Light-Curve Data:

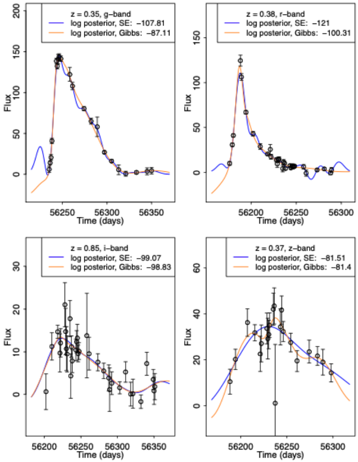


Figure 5.1: LC examples with GP fit.

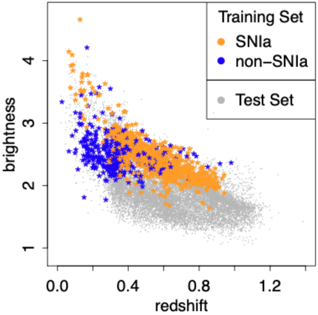


Figure 5.2: Biased training and test allocation.

STACCATO - Bias effect on classification performance:

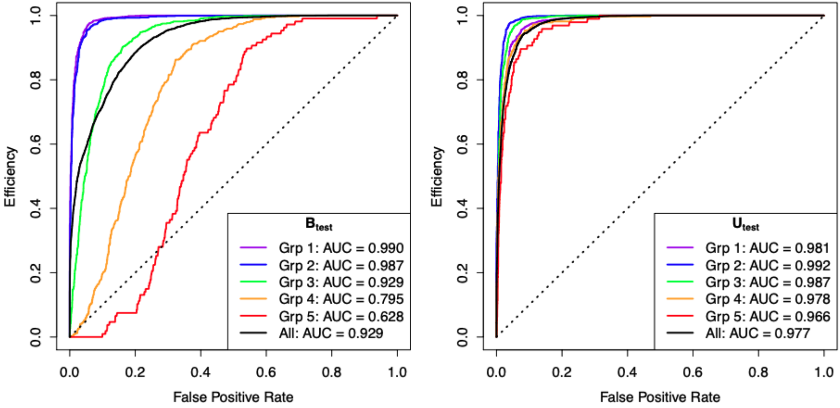


Figure 5.3: Revsbech et al. (2018)
Left: Classification performance on the spectroscopially biased training set.
Right: Randomly sampled unbiased training set (”Gold standard”).

STACCATO – Augmentation and Stratification based on Propensity Scores:

1. Propensity score $PS = P(s = 1|x)$, where $s = 1$ indicates training set assignment of sample x.
2. PS is computed with logistic regression with predictive covariates redshift and brightness.
3. Divide the data set into 5 equally-sized groups, ordered by the PS.
4. Augment the training groups with synthetic LCs sampled under the GP fit + add other training groups.
5. Compute diffusion maps for each of the training groups, including the Nyström extensions (Richards et al. 2012).
6. Train separate random forest classifier on the training and predict the test groups, respectively.

Stratification based on estimated Propensity scores:

Group	Set	Number of SNe	Number of SNIa	Proportion of SNIa
1	Training	947	652	0.69
	Test	2519	1242	0.49
2	Training	245	181	0.74
	Test	3221	1147	0.36
3	Training	17	12	0.71
	Test	3449	754	0.22
4	Training	6	6	1
	Test	3460	342	0.10
5	Training	2	0	0
	Test	3464	107	0.03

Figure 5.4: Composition of the five groups based on the estimated propensity scores. (Revsbech et al. 2018)

Optimal Training Configuration:

Test Group	Optimal training group configuration 1	2	3	4	5	AUC with synthetic LCs	AUC w/o synthetic LCs	AUC original
1	+ (0)	–	–	–	–	0.991	0.991	0.993
2	+ (0)	+ (2)	–	–	–	0.989	0.990	0.988
3	–	+ (0)	+ (0)	+ (5)	+ (5)	0.968	0.958	0.926
4	–	+ (0)	+ (5)	+ (10)	+ (5)	0.919	0.887	0.791
5	–	+ (0)	+ (6)	+ (10)	+ (2)	0.842	0.709	0.636

Figure 5.5: Optimal strata and augmentation configuration. (Revsbech et al. 2018)

- Optimal AUC of **0.961** is achieved with syntetical light curve augmentation (biased AUC: 0.929).
- 1500 test set samples used from each strata to evaluate optimal augmentation and strata configuration.

Extended STACCATO:

1. Estimate the propensity score $PS = P(s = 1|x)$, including redshift, brightness and **diffusion map** coordinates.
2. Divide the data set into 5 equally-sized groups, ordered by the PS.
3. Check **covariate balance** in related training and test strata

$$\mathrm{SMD} = \frac{(\bar{x}_{training} - \bar{x}_{test})}{\sqrt{\frac{s^2_{training} + s^2_{test}}{2}}}.$$

4. Compute diffusion maps for each of the training groups, including the Nyström extensions.
5. Train separate random forest classifier on the training and predict the test groups, respectively.

Strata comparison:

Group	Set	Number of SNe	Number of SNIa	Proportion of SNIa
1	Training	947	652	0.69
	Test	2519	1242	0.49
2	Training	245	181	0.74
	Test	3221	1147	0.36
3	Training	17	12	0.71
	Test	3449	754	0.22
4	Training	6	6	1
	Test	3460	342	0.10
5	Training	2	0	0
	Test	3464	107	0.03

Figure 5.6: Composition of the five groups based on the estimated PS. (Revsbech et al. 2018)

Strata	Set	Number of SNe	Number of SNIa	Proportion of SNIa
1	Training	996	794	0.8
	Test	2470	1759	0.71
2	Training	210	56	0.27
	Test	3256	1010	0.31
3	Training	9	0	0
	Test	3457	385	0.11
4	Training	2	1	0.5
	Test	3464	258	0.07
5	Training	0	0	NA
	Test	3466	180	0.05

Figure 5.7: Extended STACCATO: Composition of the five groups, including the diffusion map into the PS estimation.

Balance assessment:

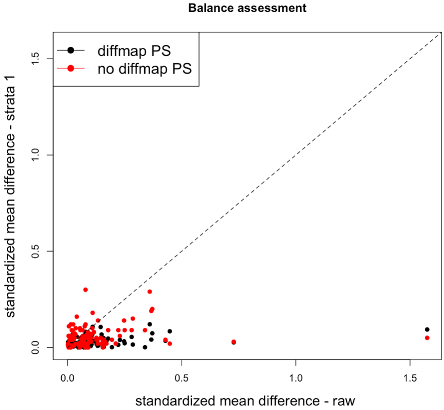


Figure 5.8: SMD between training and test data of strata 1 plotted against raw data SMD for both PS approaches.

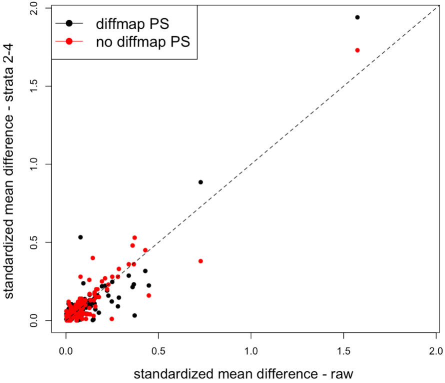


Figure 5.9: SMD between training and test data of strata 2-5 combined, plotted against raw data SMD.

Propensity score groups: Redshift vs. Brightness

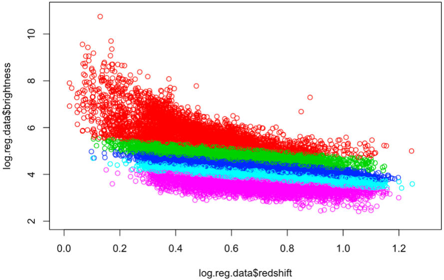


Figure 5.10: Propensity score groups using the old PS including training and test samples.

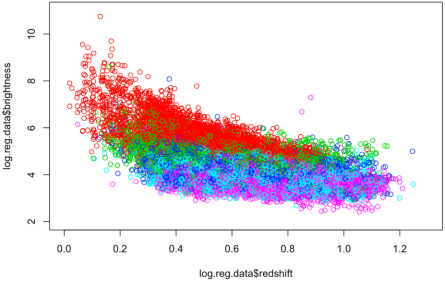


Figure 5.11: Propensity score groups using the new diffusion map PS including training and test samples.

Performance Comparison: STACCATO vs. Extended:

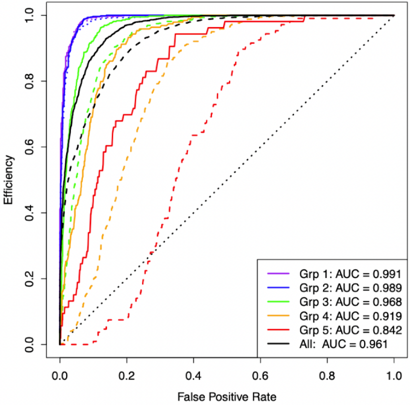


Figure 5.12: ROC curves of best STACCATO combination (Revsbech et al. 2018) using the optimized synthetical light curve augmentation.

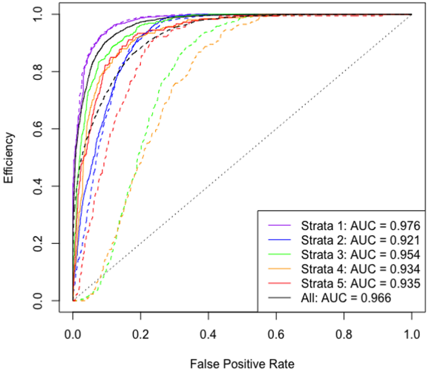


Figure 5.13: ROC curves of 'extended STACCATO' using the diffusion map included PS.

Best Performance: Extended STACCATO with redshift:

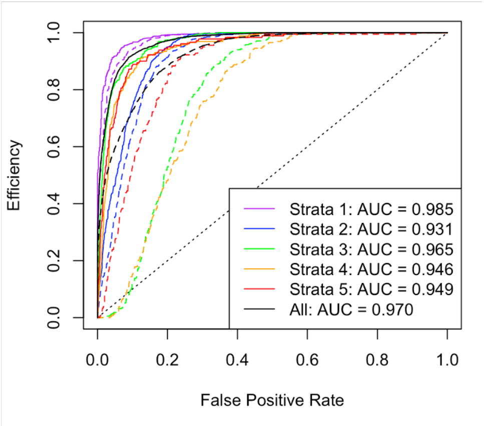


Figure 5.14: ROC curves of 'extended STACCATO' using the diffusion map included PS and redshift as predictor variable in random forest.

Updated SPCC data:

Data: Updated SPCC challenge Kessler et al. (2010a).

- 21,318 simulated supernovae of type Ia, Ib, Ic and II.
- For each SN, light curve observations are given in four color bands C= (g,r,i,z).
- Training set: 1,102 spectroscopically confirmed SNe with known types
- Test set: 20,216 SNe with unknown types and photometric information alone

Classification on updated SPCC set is more difficult due to bug-fixes in the simulations.

Updated SPCC: Strata comparison

Strata	Set	Number of SNe	Number of SNIa	Proportion of SNIa
1	Training	924	414	0.45
	Test	3340	1125	0.34
2	Training	153	125	0.82
	Test	4111	973	0.24
3	Training	21	16	0.76
	Test	4242	966	0.23
4	Training	3	2	0.67
	Test	4261	949	0.22
5	Training	1	1	1
	Test	4262	516	0.12

Figure 5.15: Old STACCATO: Composition of the five groups, including redshift and brightness into the PS estimation.

Strata	Set	Number of SNe	Number of SNIa	Proportion of SNIa
1	Training	958	518	0.54
	Test	3306	1790	0.54
2	Training	120	28	0.23
	Test	4144	927	0.22
3	Training	13	4	0.31
	Test	4250	540	0.13
4	Training	7	4	0.57
	Test	4261	949	0.22
5	Training	4	4	1
	Test	4259	662	0.16

Figure 5.16: Extended STACCATO: Composition of the five groups, including the diffusion map into the propensity score estimation.

Updated SPCC – Balance Assessment:

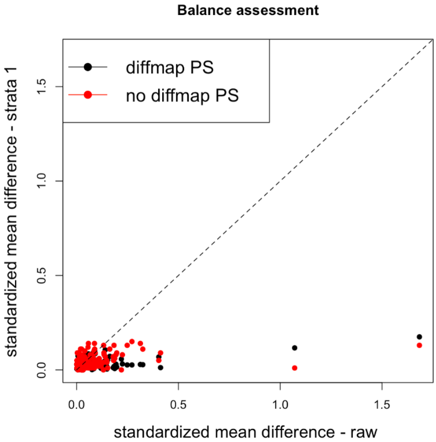


Figure 5.17: SMD between training and test data of stratum 1.

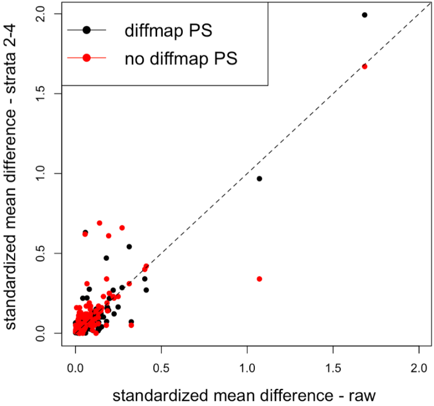


Figure 5.18: SMD between training and test data of strata 2-5.

Extended STACCATO results on updated SPCC

	AUC bias	Comp. 1	Comp. 2	Comp. 2 + red
grp1	0.984	0.981	0.981	0.988
grp2	0.890	0.906	0.959	0.966
grp3	0.780	0.952	0.954	0.959
grp4	0.848	0.950	0.951	0.955
grp5	0.910	0.939	0.938	0.943
all	0.902	0.944	0.953	**0.955**

Table 1: AUC results of extended STACCATO on updated SPCC data. No synthetical data augmentation. Different training strata compositions compared.

- 'Gold standard' on unbiased (randomly selected) set: 0.961.

State-of-the-art on updated SPCC data:

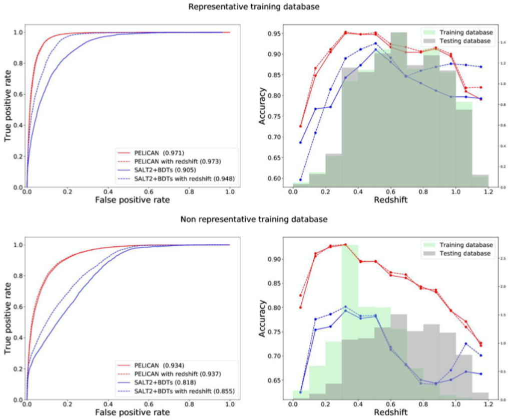


Figure 5.19: (Pasquet et al. 2019)

Results by Lochner et al. (2016); Pasquet et al. (2019) on the updated SPCC data. Best performance: AUC 0.934.

Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) (Kessler et al. 2019)

ID[a]	Object type	$N_{\text{confirmed}}$	$N_{\text{unconfirmed}}$	Galactic	Weight[b]
90	Type Ia SN	2,313	1,659,831	No	1
67	Peculiar Type Ia SN – 91bg-like	208	40,193	No	1
52	Peculiar Type Ia SN – SNIax	183	63,664	No	1
42	Type II SN	1,193	1,000,150	No	1
62	Type Ibc SN	484	175,094	No	1
95	Superluminous SN (Magnetar model)	175	35,782	No	1
15	Tidal disruption event	495	13,555	No	2
64	Kilonova	100	131	No	2
88	Active galactic nuclei	370	101,424	No	1
92	RR Lyrae	239	197,155	Yes	1
65	M-dwarf stellar flare	981	93,494	Yes	1
16	Eclipsing binary stars	924	96,472	Yes	1
53	Mira variables	30	1,453	Yes	1
6	Microlens from single lens	151	1,303	Yes	1
991[c]	Microlens from binary lens	0	533	Yes	2
992[c]	Intermediate luminous optical transient	0	1,702	No	2
993[c]	Calcium rich transient	0	9,680	No	2
994[c]	Pair instability SN	0	1,172	No	2
	Total	7,846	3,492,888		

Figure 5.20: PLAsTiCC data summary. (Boone 2019; Kessler et al. 2019)

State-of-the-art SN classification (Boone 2019):

Metric name	Flat-weighted classifier	Redshift-weighted classifier
Flat-weighted metric	0.468	0.510
Redshift-weighted metric	0.523	0.500
Kaggle metric	0.649	0.709
AUC – 90: Type Ia SN	0.95721	0.95204
AUC – 67: Peculiar Type Ia SN – 91bg-like	0.96672	0.96015
AUC – 52: Peculiar Type Ia SN – SNIax	0.85988	0.84203
AUC – 42: Type II SN	0.93570	0.90826
AUC – 62: Type Ibc SN	0.92851	0.91558
AUC – 95: Superluminous SN (Magnetar model)	0.99442	0.99257
AUC – 15: Tidal disruption event	0.99254	0.99243
AUC – 64: Kilonova	0.99815	0.99579
AUC – 88: Active galactic nuclei	0.99772	0.99706
AUC – 92: RR Lyrae	0.99987	0.99986
AUC – 65: M-dwarf stellar flare	0.99999	0.99999
AUC – 16: Eclipsing binary stars	0.99983	0.99983
AUC – 53: Mira variables	0.99947	0.99937
AUC – 6: Microlens from single lens	0.99962	0.99966

Figure 5.21: Best classification results during blinded (PLAsTiCC) challenge (Boone 2019).

STACCATO - summary

- State-of-the-art results on SPCC data without data augmentation.
- Strata selection has to be validated (e.g. stratified or weighted cross-validation).
- Applying STACCATO on PLAsTiCC data set (Boone 2019)

- Generalization of methodology
- Application of covariate shift and domain adaptation methods in astronomy

References I

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Bickel, S., Brückner, M., and Scheffer, T. (2009). Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155.

Bickel, S. and Scheffer, T. (2007). Dirichlet-enhanced spam filtering based on biased samples. In *Advances in neural information processing systems*, pages 161–168.

Boone, K. (2019). Avocado: Photometric classification of astronomical transients with gaussian process augmentation. *The Astronomical Journal*, 158(6):257.

Cortes, C., Mansour, Y., and Mohri, M. (2010). Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450.

References II

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Daumé III, H. (2009). Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Daume III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126.

Ganin, Y. and Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). Technical report, National Bureau of Economic Research.

References III

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul):1391–1445.

Katole, A. L., Yellapragada, K. P., Bedi, A. K., Kalra, S. S., and Chaitanya, M. S. (2015). Hierarchical deep learning architecture for 10k objects classification. *arXiv preprint arXiv:1509.01951*.

Kessler, R., Bassett, B., Belov, P., Bhatnagar, V., Campbell, H., Conley, A., Frieman, J. A., Glazov, A., González-Gaitán, S., Hlozek, R., et al. (2010a). Results from the supernova photometric classification challenge. *Publications of the Astronomical Society of the Pacific*, 122(898):1415.

References IV

Kessler, R., Conley, A., Jha, S., and Kuhlmann, S. (2010b). Supernova photometric classification challenge. *arXiv preprint arXiv:1001.5210*.

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P., Chernoff, D., Connolly, A., Dai, M., Daniel, S., et al. (2019). Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501.

Kouw, W. M. and Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Lochner, M., McEwen, J. D., Peiris, H. V., Lahav, O., and Winter, M. K. (2016). Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*, 225(2):31.

References V

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Mahmud, M. H. (2009). On universal transfer learning. *Theoretical Computer Science*, 410(19):1826–1846.

Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

References VI

Pasquet, J., Pasquet, J., Chaumont, M., and Fouchez, D. (2019). Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, 627:A21.

Revsbech, E. A., Trotta, R., and van Dyk, D. A. (2018). Staccato: a novel solution to supernova photometric classification with biased training sets. *Monthly Notices of the Royal Astronomical Society*, 473(3):3969–3986.

Richards, J. W., Homrighausen, D., Freeman, P. E., Schafer, C. M., and Poznanski, D. (2012). Semi-supervised learning for photometric supernova classification. *Monthly Notices of the Royal Astronomical Society*, 419(2):1121–1135.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

References VII

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Sugiyama, M., Krauledat, M., and MÃžller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155.

References VIII

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.

Umer, M., Frederickson, C., and Polikar, R. (2019). Vulnerability of covariate shift adaptation against malicious poisoning attacks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

References IX

Vidyasagar, M. (2002). *A theory of learning and generalization*. Springer-Verlag.

Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827.

Thank you very much for your time!