Transfer and Active Learning for Efficient Machine Learning

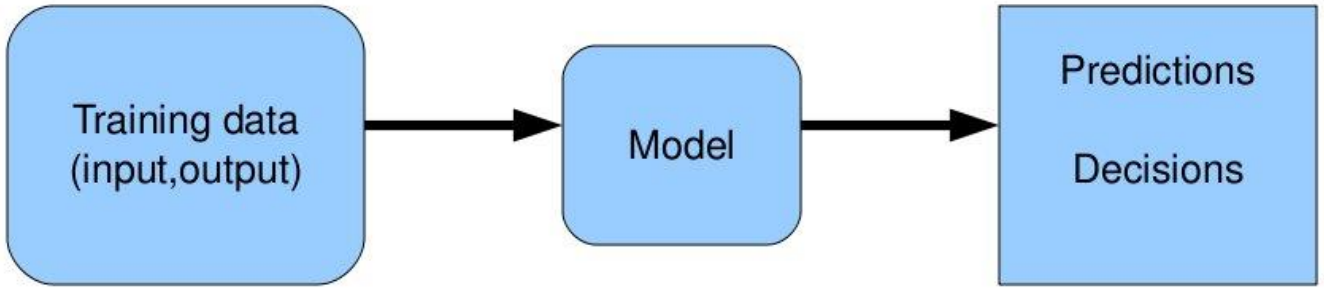
Adriana Bîrluţiu
1 Decembrie 1918
University of Alba Iulia
adriana.birlutiu@uab.ro

Outline

- Transfer Learning
  - Approaches
  - TL application in chemometrics
  - Transfer component analysis (TCA)
  - Multi-TCA
  - Experimental evaluation
- Active Learning
  - Active Transfer criterion
  - Application in a project for defect detection in porcelainware

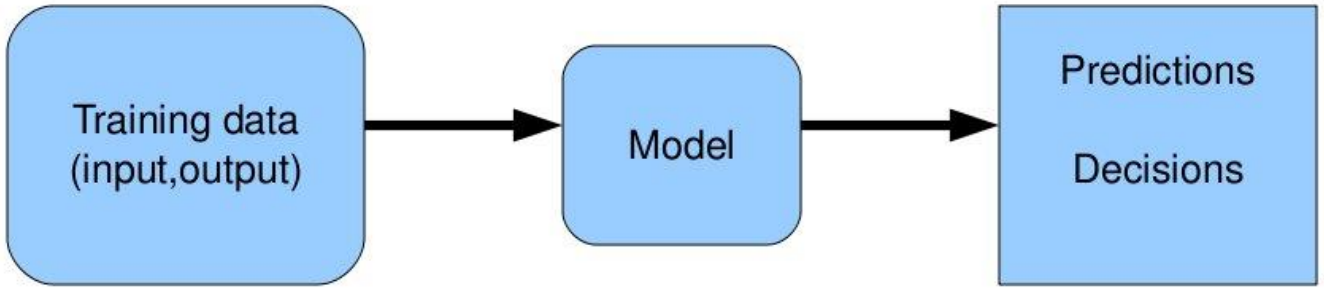
Machine Learning

- Branch of AI focused on the design and development of methods that allow machines to learn based on observations



Machine Learning

- Branch of AI focused on the design and development of methods that allow machines to learn based on observations



- Obtaining labeled data to train the algorithms is expensive!

Efficient machine learning

Characteristics of (human) learning:

- Based on prior experience
  - transfer learning (e.g. C++ -> Java)
- Selects the most useful information
  - active learning (selects the most useful information)

Transfer learning

- Fundamental assumption in machine learning:
  - Data is i.i.d. (independent and identically distributed)
  - Training and test data stem from the same distribution
- Often, this assumption does not hold
- Transfer learning addresses the mismatch between training and test data

Traditional vs transfer learning

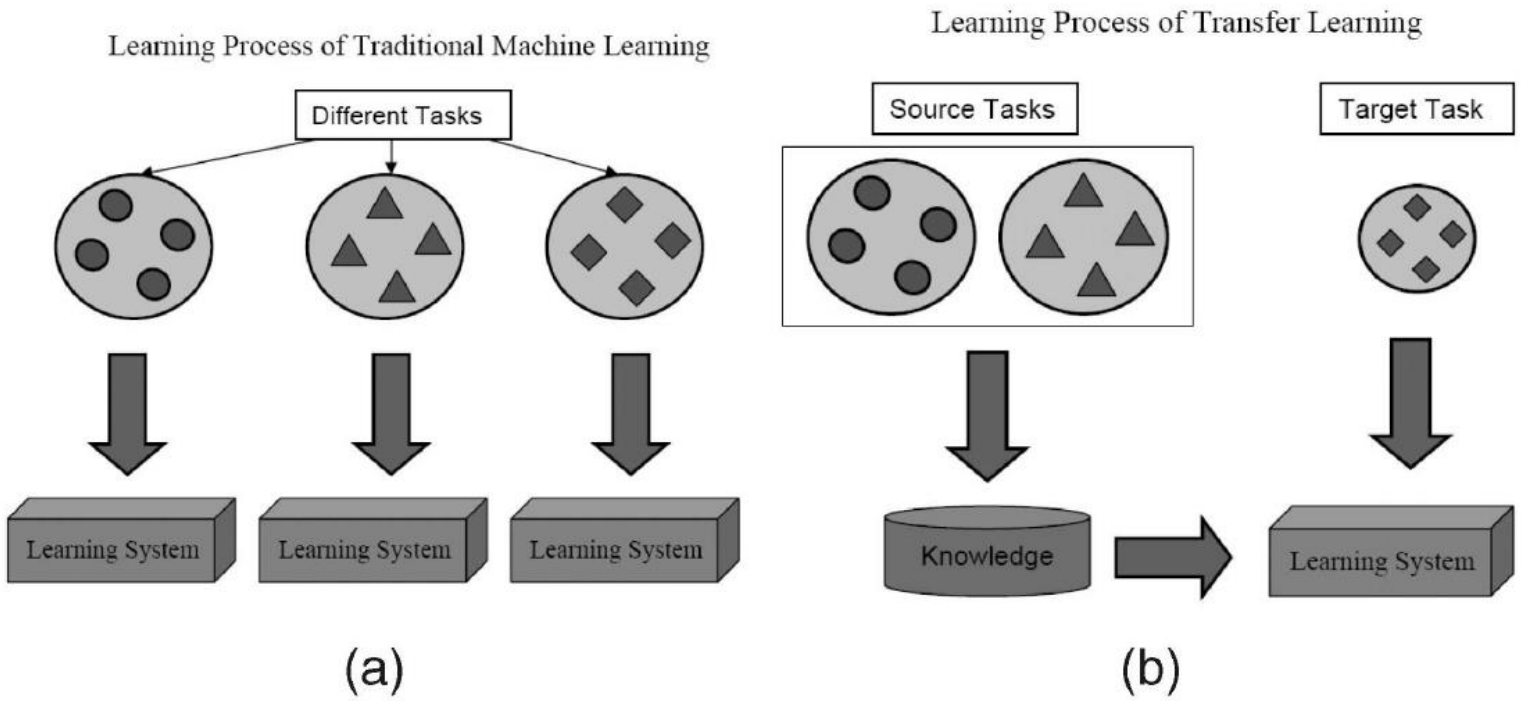


Fig. 1. Different learning processes between (a) traditional machine learning and (b) transfer learning.

(adapted from Pan et al 2010, A survey on Transfer Learning)

Approaches

- Instance-based: reweighted source data are used for learning in the target space
- Parameter-based: source and target model share some common parameters or a prior distribution (e.g. Hierarchical Learning)
- Feature-based: source knowledge is used for learning a good feature representation in the target space (e.g. Transfer Component Analysis)

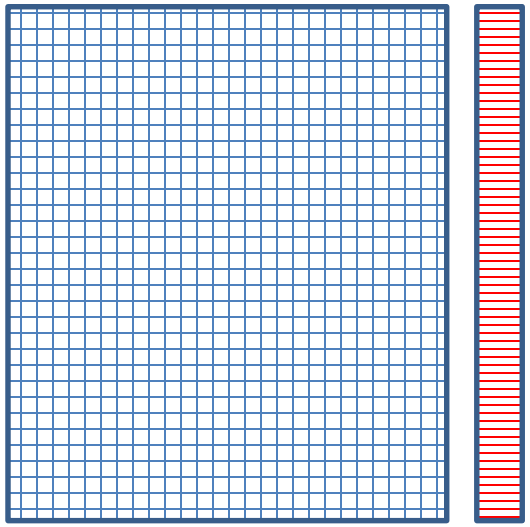
TL for chemometric application

- Application: control the polymerization process of melamine based on spectroscopic data (NIR), measured in-line at an industrial partner (DYNEA)
- Regression problem: predict the temperature of a sample based on spectroscopic data
- Transfer learning settings:
  1. Change of lamp
  2. Change in the recipe of the composition
  3. …
- as very often:
  - Spectra easy to obtain
  - Reference values cumbersome / expensive to measure

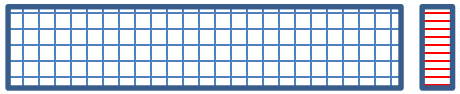
B. Malli, A. Birlutiu, T. Natschlaeger. ***Standard-free calibration transfer-An evaluation of different techniques.*** *Chemometrics and Intelligent Laboratory Systems, vol. 161, pp. 49-60, 2017*

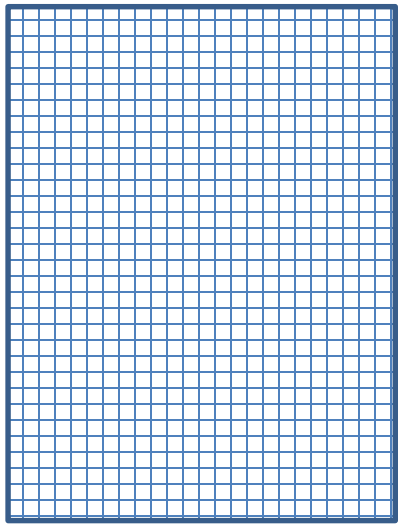
Data

Source data:
spectra + reference values



Target data:
A few spectra + reference values
A lot of spectra without reference values
(unlabelled data)





Question:
How do we combine these data?

Approach

- Use only target labelled data and ignore any other source data (no transfer)

Approach

- Use only target labelled data and ignore any other source data (no transfer)
- Use source data + target labelled data (all labelled data pulled together)

Approach

- Use only target labelled data and ignore any other source data (no transfer)
- Use source data + target labelled data (all labelled data pulled together)
- Combine all data using a more "sophisticated" method

Transfer component analysis (TCA)

- Source domain (S), target domain (T)
- Assumption: $\mathrm{P}(\mathrm{X_S}) \neq P(X_T)$
  - holds for the chemometric application since the conditions under which the spectra were obtained are different between domains

- Intuition: discover a good feature representation across domains
- Idea: maps data in a shared subspace s.t.
  - distance between distributions is minimized
  - data properties are preserved

- Goal: find a feature map $\phi: X \to H$ where H is a RKHS such that $P(\phi(X_S)) \approx P(\phi(X_T))$ constraints: variance of data is preserved
- Key assumption:
  $P(X_S) \neq P(X_T)$ but $P(Y_S|\phi(X_S)) = P(Y_T|\phi(X_T))$

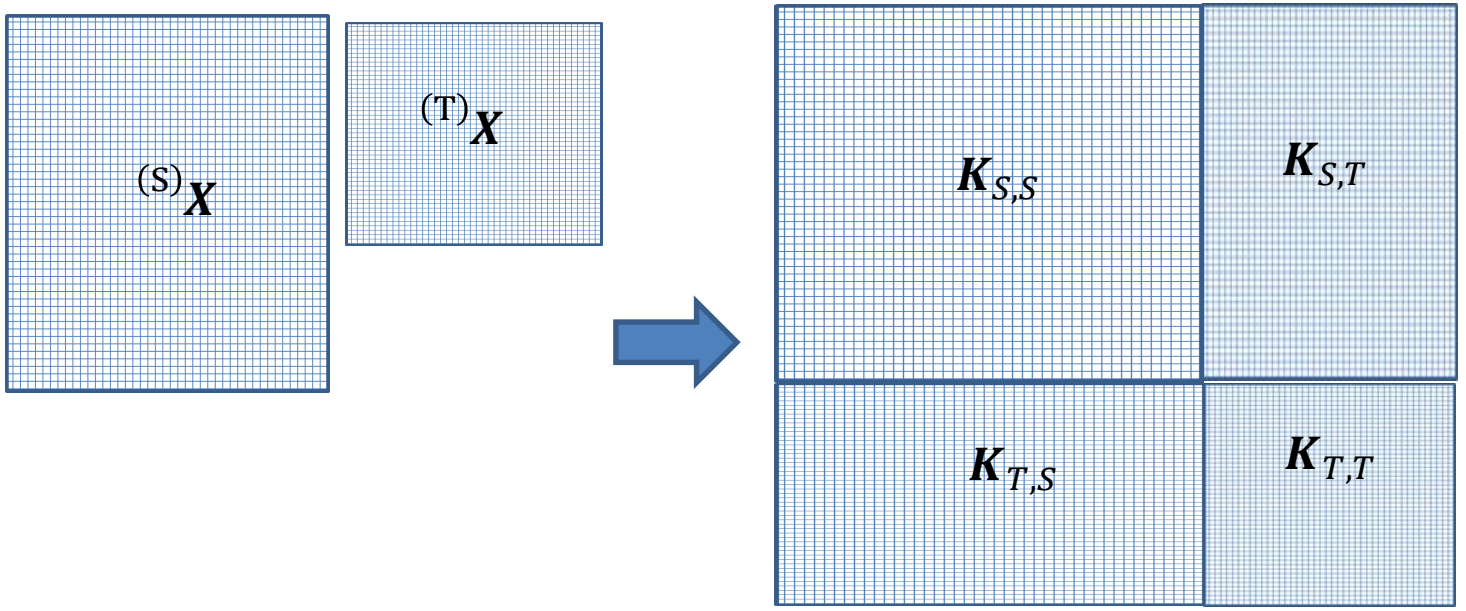
[Pan et. al, 2011, IEEE TNN]

Distance between distributions

- Distance between distributions, e.g. Kullback-Leibler divergence
- Maximum Mean Discrepancy [Gretton et. al 2007]: distance between distributions = distance between the means of the two samples mapped in a RKHS
- Let $X_s^{'} = \{\phi(X_{S_i})\}$ and $X_T^{'} = \{\phi(X_{T_i})\}$

$$\text{Dist}(X_S', X_T') = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_{S_i}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_{T_i}) \right\|_{\mathcal{H}}^2 .$$

Use kernels for finding ϕ



Kernel trick: $k(x_i, x_j) = \phi(x_i)'\phi(x_j)$

$$\text{Dist}(X'_S, X'_T) = \text{tr}(KL),$$

$$K = \begin{bmatrix} K_{S,S} & K_{S,T} \\ K_{T,S} & K_{T,T} \end{bmatrix}$$

and $L = [L_{ij}] \succeq 0$ with $L_{ij} = \frac{1}{n_1^2}$ if $x_i, x_j \in X_S$; $L_{ij} = \frac{1}{n_2^2}$ if $x_i, x_j \in X_T$; otherwise, $-\frac{1}{n_1 n_2}$.

TCA Approch

- First step: apply TCA to learn a low-dimensional subspace:
  1. the distributions of the source and target domain data are close to each other
  2. data properties (such as variance) are preserved

- Second step: apply a regressor in the shared TCA subspace to train regression models across domains (use the projected source and target labelled data)

Multi-TCA

- Multi-TCA an extension of TCA to domain generalization
- Multiple source and target domains
- Domain generalization: no input data from target domains but the characteristics of target data are sufficiently captured by X1, X2,...XS
- Goal: find $\phi: X \to H$ a feature map and H a RKHS

$P(\phi(X_1)) \approx \cdots \approx P(\phi(X_S))$ under some constraints

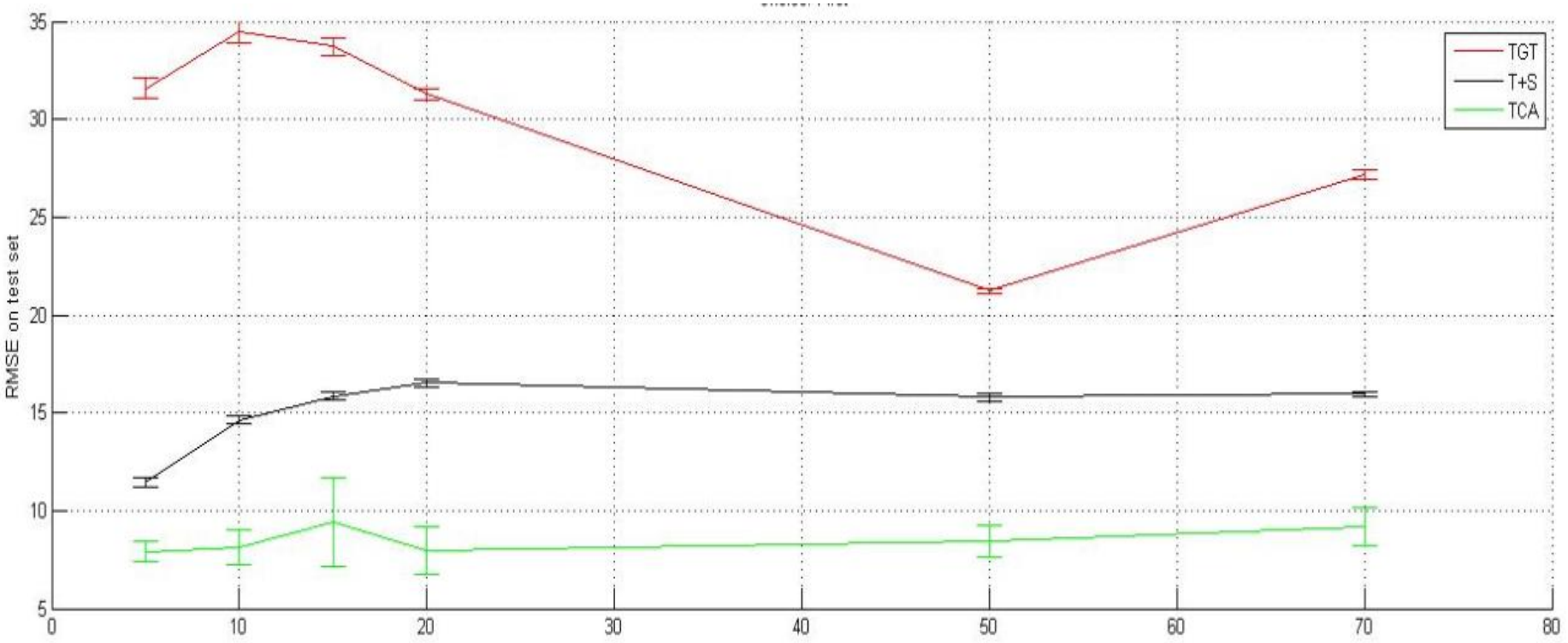
- *T. Grubinger, A. Birlutiu, H. Schoner, T. Natschlager, T. Heskes.* ***Multi-Domain Transfer Component Analysis for Domain Generalization.*** *Neural Processing Letters, 2017.*
- T. Grubinger, A. Birlutiu, H. Schoner, T. Natschlager, T. Heskes. **Domain Generalization based on Transfer Component Analysis.** IWANN, 2015

Performance comparison

- 48 learning settings
- 12 TL methods (simple, instance weighting, calibration transfer, instrument standardization in chemometrics) + TCA
- 3 choice methods of labelled data (First, Random, KennardStone)

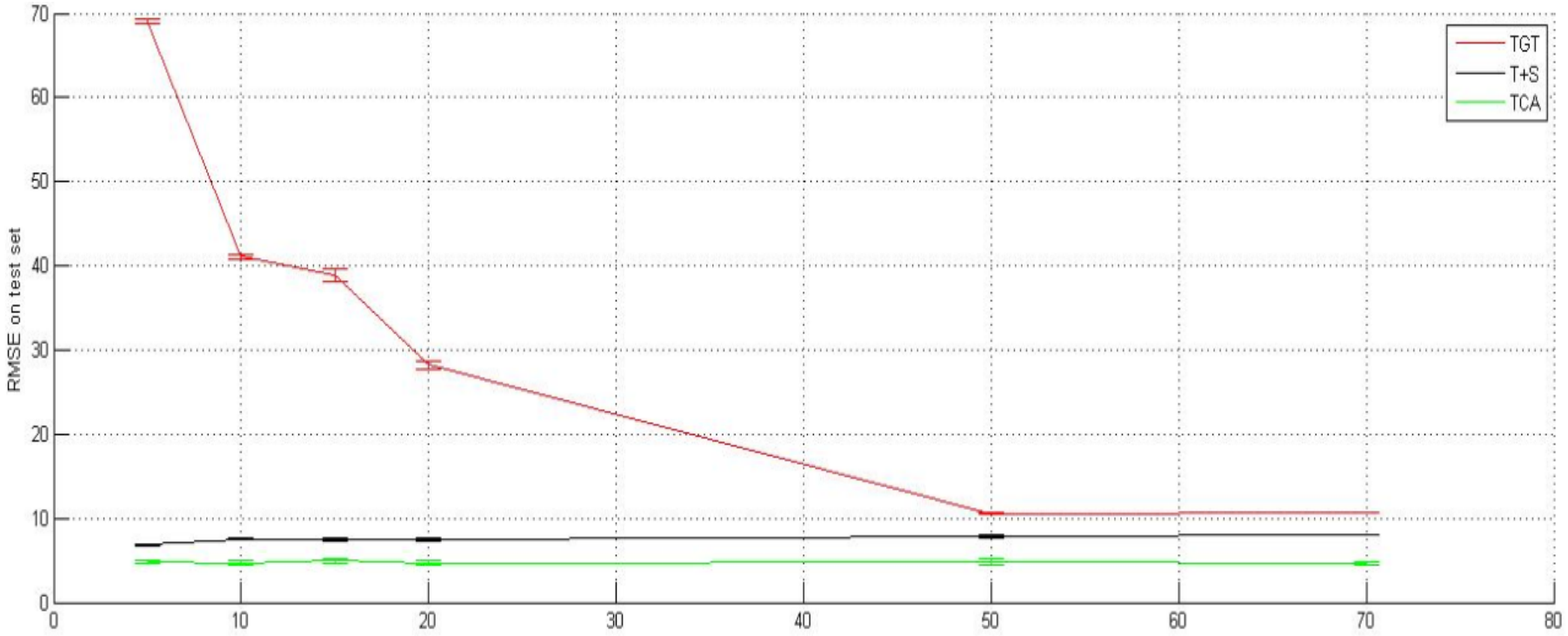
- Use only target labelled data and ignore any other source data (no transfe: TGT in the next figure)
- Use source data + target labelled data (all labelled data pulled together: T+S in the next figure)
- Combine all data using a more "sophisticated" method (TCA in the next figure)

Experimental evaluation



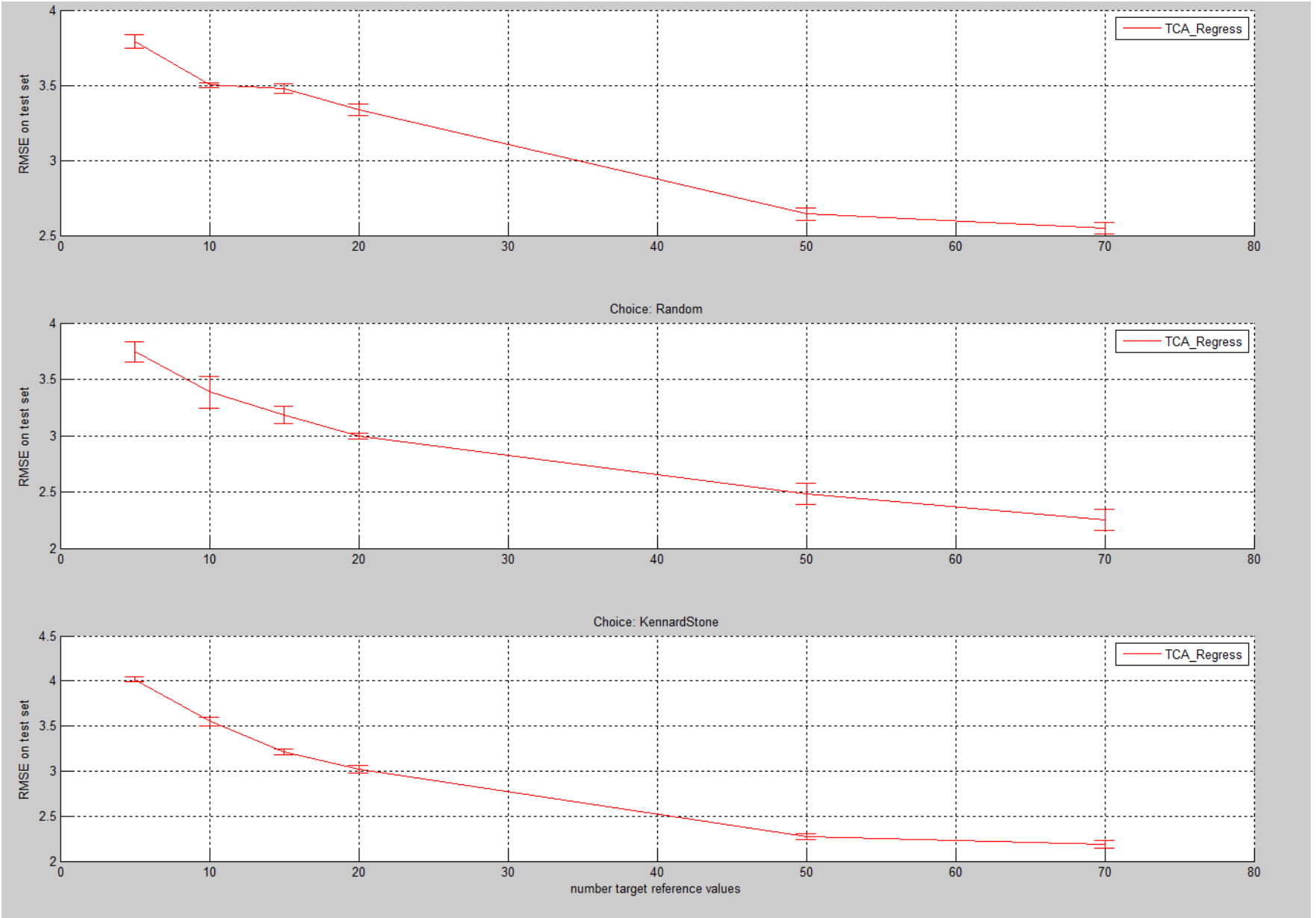
Difference between source and target domains: Change of reactor

Experimental evaluation



Difference between source and target domains: Change of lamp

RMSE decreases with more labelled data. Choice of the labelled points



Experimental evaluation for TCA applied in chemometrics

Conclusions

- TCA is robust: it has among the best performance in most of the learning settings
- TCA can be further optimized:
  - wider parameter range when selecting the optimal parameters
  - use of non-linear kernels
  - STCA: an extension of TCA that takes into account labels in the source domain and the manifold information (computational more expensive than TCA)
  - Extension to multiple source domains

Some Research Issues when applying TL techniques

- Given a specific application, which kind of transfer learning methods should be used?
- How to avoid negative transfer? Given a target domain/task, how to find source domains/tasks to ensure positive transfer
- Transfer learning meets active learning => Efficient machine learning

Transfer and Active Learning

1. Can we use the data from similar tasks when learning a current a task?
2. Can we learn faster by optimally selecting the data points to label?

Active Learning Criteria

- Uncertainty sampling criterion: chooses for labeling the example for which the model's predictions are most uncertain.The uncertainty of the predictions can be measured, for example, using Shannon entropy

$$\text{Uncertainty(x)} = -\sum_{y} p(y|x) \log p(y|x).$$

- Variance reduction criterion: the accuracy with which the parameters of the model can be estimated. In the Bayesian context ⇔ reduction in the entropy of the posterior distribution over model parameters
- Expected model chance: chooses as the most informative query the one which when added to the training set would yield the greatest model change. Quantifying the model change depends on the learning framework. In the Bayesian setting, the model change can be quantified via a distance measure between distributions

Active Learning based on Transfer Learning

- Query by Committee criterion: selects those points with the highest disagreement between the classification models, induced by the uncertainty in their distribution
- AT criterion exploits learning with multiple data sets and use the learned models of other data sets when determining how informative a new data point is

$$\mathrm{AT(x)} = \sum_{m=1}^{M} \frac{1}{M} \mathrm{KL}[\,\overline{p}(\cdot|x) \| p_m(\cdot|x)\,]\,,$$

- *D. Onita, A. Birlutiu.* ***Active Learning based on Transfer Learning Techniques for Image Classification*** *ESANN, 2018.*
- *A. Birlutiu, P. Groot, T. Heskes.* ***Efficiently Learning the Preferences of People.*** *Machine Learning Journal, ISSN: 0885-6125, pp. 1-28, 2013.*

Quality control in the porcelain industry

- Performed manually
- Expensive process, which requires trained personnel
- Prone to human error
- The need for automated inspection!

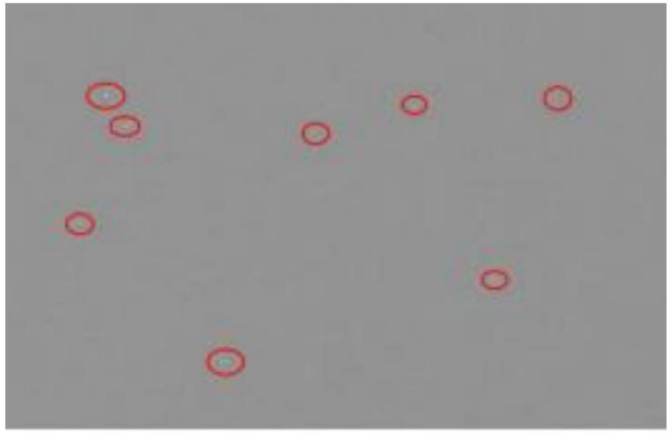
UEFISCDI/ PN-III-P2-2.1-BG-2016-0333 ***Intelligent system based on machine learning and computer vision for the optimization of the manufacturing process of porcelain (SIVAP)***

Different types of defects

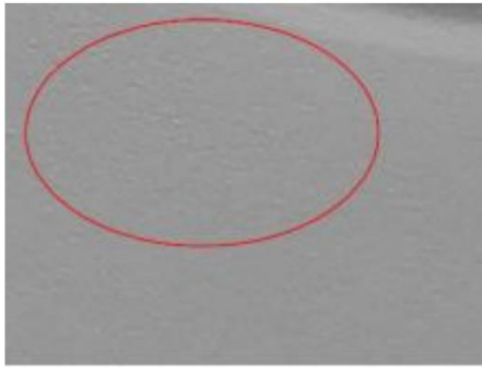
- 2D defects
- 3D defects
- structure defects

(a) Deterioration after pressing.

(d) Margin deformation.



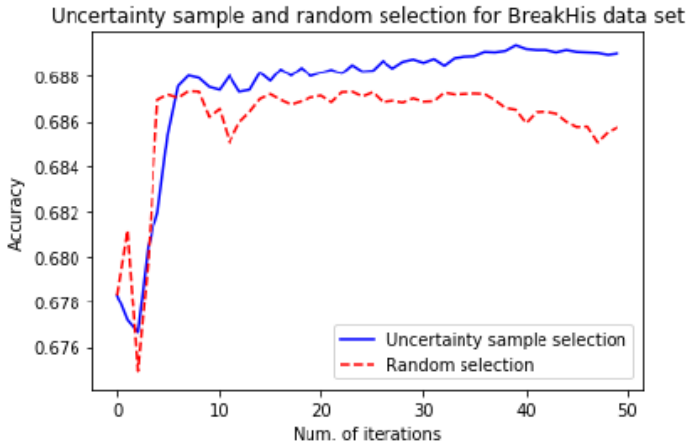
(b) Bumps.



(c) Texture defects.

Experimental evaluation

- Data set with plate images: defective and correct
- The training data was used as a pool out of which points were selected for labeling either randomly or actively.
- After selection of a point, either active or random, the point was added to the training data and deleted from unlabeled data.
- The model was retrained on the new training set and predictions were made on the validation set (50 retraining).
- Averaged results over 20 splittings of data into training, unlabeled and validation sets
- Preprocessing of images
  - Converting each image to gray scale
  - Resizing to 28x28 (784-dimensional feature vector)
  - Centering of data around zero.
  - Whitening the data
- Comparison of different learning algorithms (**SVM**, Logistic Regression, Random forest)



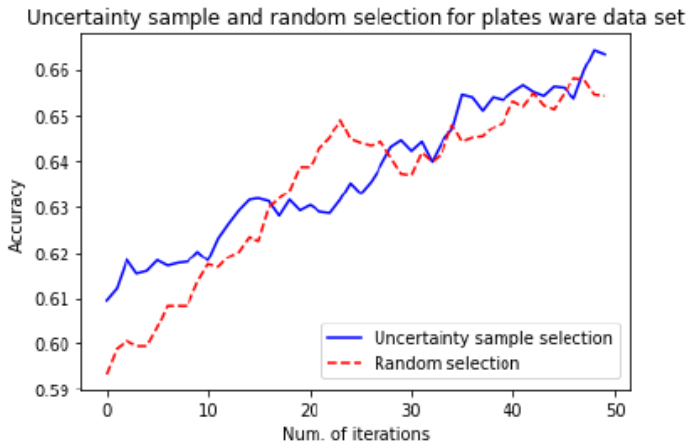
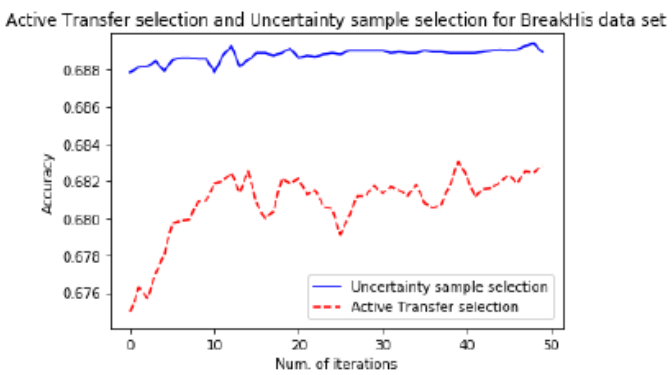


Fig. 2: Comparison of accuracies obtained with random versus active selection of training points. Uncertainty Sampling criterion was used for actively selecting training points. Left: BreakHis data set. Right: porcelain ware data set.



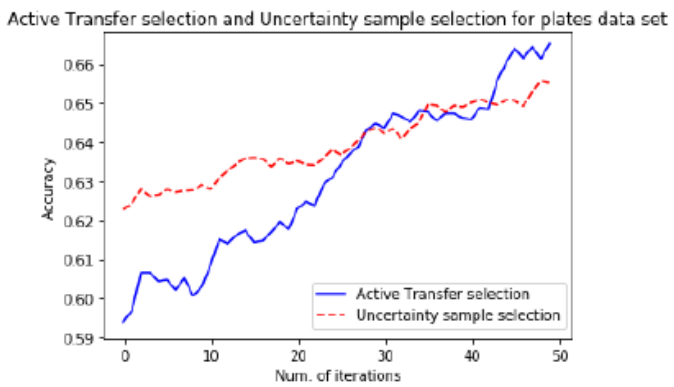


Fig. 3: Comparison of accuracies obtained with two strategies of actively selecting training points: Active Transfer criterion and Uncertainty Sampling. Left: BreakHis data set. Right: porcelain ware data set.

Outline

- Transfer Learning
  - Approaches
  - TL application in chemometrics
  - Transfer component analysis (TCA)
  - Multi-TCA
  - Experimental evaluation
- Active Learning
  - Active Transfer criterion
  - Application in a project for defect detection in porcelainware

References

- Wang, X., Huang, T.-K., Schneider, J., Active transfer learning under model shift. Proceeding ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, Pages II-1305-II-1313
- Wang, X. Active Transfer Learning, PhD Thesis. CMU, 2016.
- Birlutiu A., Groot P., Heskes T. (2013) Eciently learning the preferences of people. Machine Learning Journal, 90 (1), pp.1-28, Springer, ISSN: 0885-6125.
- B. Malli, A. Birlutiu, T. Natschlaeger. Standard-free calibration transfer-An evaluation of different techniques.Chemometrics and Intelligent Laboratory Systems, vol. 161, pp. 49–60, 2017
- Zhao, L., Jialin Pan, S., Wei Xiang, E., Zhong,E., Lu, Z., and Yang, Q. Active transfer learning for cross-system recommendation. In Proceedings of the 27th AAAI Conference on Artificial Intelligence, 2013.
- Pan S.J., Yang Q. (2010) A survey on transfer learning. IEEE Trans. Knowle. Data Eng. 2010;vol. 22, pp. 13451359.

Thank you for your attention!