Introduction to Transfer Learning

Sergio Peignier
In collaboration with:
Mathilde Mougeot and **Nicolas Vayatis**

CMLA - ENS Cachan, Université Paris-Saclay, France

Outline

Outline

General concept

- **General concept** (machine-learning, psychology, education).
- **Query "transfer learning"** using the *Carrot*^{2} Open Source Search Results Clustering Engine.
    - Machine Learning: Domain adaptation, Pre-trained models, ...
    - Psychology/Education: Metacognition, Language, Higher education, Learning theory, ...

- **Different communities** share very **similar concepts** (positive/negative transfer, source/target domain ...).

General Definition

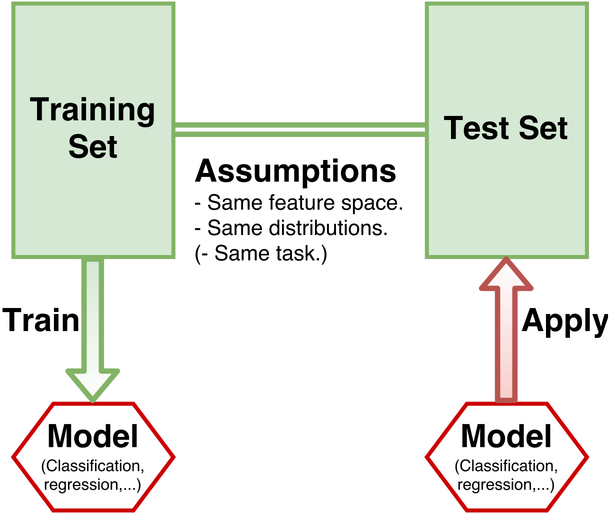
Transfer Learning: Capacity to **apply knowledge** learned in **one context** in **different ones** ([Schunk 2012] education).

- **Informal example**: Learning the piano is easier if one already plays another instrument.
- Recognise some **common elements** (features, principles, concepts ...) between a **previous** context and a **new** one.
- **Assess** the value of **applying** them in the new context.

Outline

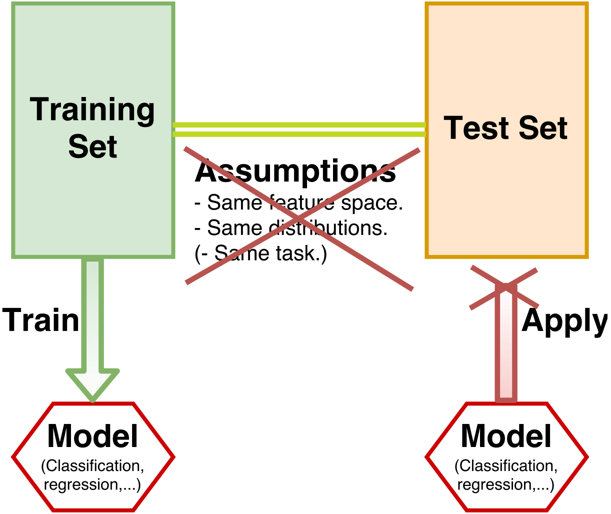
1. Introduction
   - General concept
   - Transfer Learning in Machine Learning
   - Formal definition
2. Transfer Learning Taxonomy
   - Labels availability
   - Domain differences vs task differences
   - Feature space differences
   - Methodology (what is being transferred)
3. Related tasks
   - Self-taught learning
   - Multi-task learning
   - Adaptation of probability distributions
4. Negative Transfer

Traditional scenario.



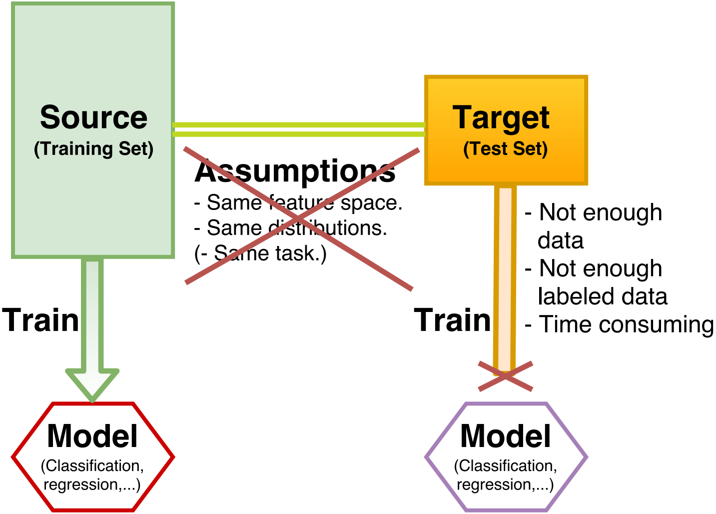
Transfer Learning in Machine Learning

$\neq$ feature space or $\neq$ distributions or $\neq$ task **but related**

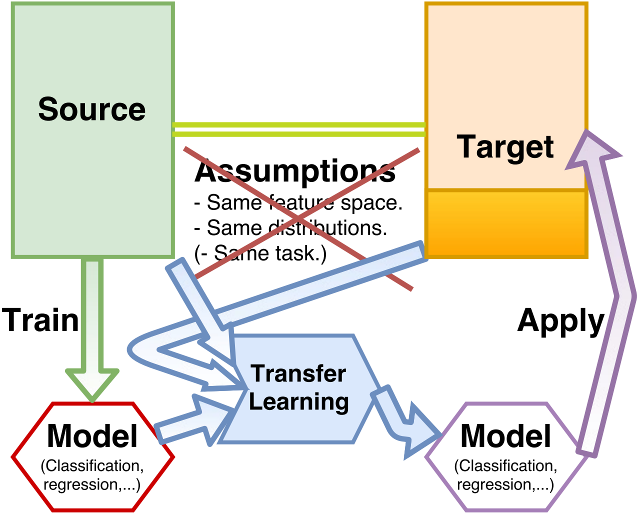


Transfer Learning in Machine Learning

Training a new model is not possible



Transfer Learning: Overcome the domain difference



Outline

1. Introduction
   - General concept
   - Transfer Learning in Machine Learning
   - Formal definition
2. Transfer Learning Taxonomy
   - Labels availability
   - Domain differences vs task differences
   - Feature space differences
   - Methodology (what is being transferred)
3. Related tasks
   - Self-taught learning
   - Multi-task learning
   - Adaptation of probability distributions
4. Negative Transfer

Domain and Task

- **Domain** $\mathcal{D} = \{\mathcal{X}, P(X)\}$
  - $\mathcal{X}$: **Feature space** s.t. **Dataset** $X = \{x_1, x_2, \dots\} \in \mathcal{X}$
  - $P(X)$: **Probability distribution** of X
- **Task** $\mathcal{T} = \{\mathcal{Y}, f(.)\}$
  - $\mathcal{Y}$: **Label space**
  - $f(.)$: **Predictive function**
    learnt from pairs $\{x_i, y_i\}$ (Probabilistically: $P(y|x)$)

- **Source** $\mathcal{D}_S = \{\mathcal{X}_S, P_S(X)\}$, $\mathcal{T}_S = \{\mathcal{Y}_S, f_S(.)\}$
  Source dataset: $D_S = \{(x_{S1}, y_{S1}), (x_{S2}, y_{S3}), \dots\}$
- **Traget** $\mathcal{D}_T = \{\mathcal{X}_T, P_T(X)\}$, $\mathcal{T}_T = \{\mathcal{Y}_T, f_T(.)\}$
  Traget dataset: $D_T = \{(x_{T1}, y_{T1}), (x_{T2}, y_{T3}), \dots\}$

Transfer Learning (TL) Definition [Weiss et al. 2016]

"Process of **improving the target predictive function** $f_T(.)$ by using the **related information** from $\mathcal{D}_S$ and $\mathcal{T}_S$" [Weiss et al. 2016]

Where: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$

- If $\mathcal{D}_S \neq \mathcal{D}_T$:
  - $\mathcal{X}_S \neq \mathcal{X}_T$: **Heterogenous TL**
  - $\mathcal{X}_S = \mathcal{X}_T$ **but** $P(X_S) \neq P(X_T)$: **Homogenous TL**
- If $\mathcal{T}_S \neq \mathcal{T}_T$:
  - $\mathcal{Y}_S \neq \mathcal{Y}_T$: **Mismatch in label space**
  - $\mathcal{Y}_S = \mathcal{Y}_T$ **but** $f_S(.) \neq f_T(.)$ **Different predictive function** ($P(Y_S|X_S) \neq P(Y_T|X_T)$ conditional probability distributions)

- Single source domain definition $\rightarrow$ **multiple source domain**.

Outline

Labels availability

"Process of **improving the target predictive function** $f_T(.)$ by using the **related information** from $\mathcal{D}_S$ and $\mathcal{T}_S$" [Weiss et al. 2016]

Where: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$

- If $\mathcal{D}_S \neq \mathcal{D}_T$:
  - $\mathcal{X}_S \neq \mathcal{X}_T$: **Heterogenous TL**
  - $\mathcal{X}_S = \mathcal{X}_T$ **but** $P(X_S) \neq P(X_T)$: **Homogenous TL**
- If $\mathcal{T}_S \neq \mathcal{T}_T$:
  - $\mathcal{Y}_S \neq \mathcal{Y}_T$: **Mismatch in label space**
  - $\mathcal{Y}_S = \mathcal{Y}_T$ **but** $f_S(.) \neq f_T(.)$ **Different predictive function** ($P(Y_S|X_S) \neq P(Y_T|X_T)$ conditional probability distributions)

Labels availability

- [Daumé et al. 2007], [Chattopadhyay et al. 2011]
  - **Supervised TL**: **Many source** labels; **few target** ones.
  - **Semi-supervised TL**: **Many source** labels; **no target** ones.

Labels availability

- [Daumé et al. 2007], [Chattopadhyay et al. 2011]
  - **Supervised TL**: **Many source** labels; **few target** ones.
  - **Semi-supervised TL**: **Many source** labels; **no target** ones.
- [Gong et al. 2012], [Blitzer et al. 2006]
  - **Semi-supervised TL**: **Many source** labels; **few target** ones.
  - **Unsupervised TL**: **Many source** labels; **no target** ones.

Labels availability

- [Daumé et al. 2007], [Chattopadhyay et al. 2011]
  - **Supervised TL**: **Many source** labels; **few target** ones.
  - **Semi-supervised TL**: **Many source** labels; **no target** ones.
- [Gong et al. 2012], [Blitzer et al. 2006]
  - **Semi-supervised TL**: **Many source** labels; **few target** ones.
  - **Unsupervised TL**: **Many source** labels; **no target** ones.
- [Cook et al. 2012], [Feuz et al. 2014]
  - **Source labels**:
    Present → **supervised**; Absent → **unsupervised**.
  - **Target labels**:
    Present → **informed**; Absent → **uninformed**.

Outline

Domain differences vs tasks differences

"Process of **improving the target predictive function** $f_T(.)$ by using the **related information** from $\mathcal{D}_S$ and $\mathcal{T}_S$" [Weiss et al. 2016]

Where: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$

- If $\mathcal{D}_S \neq \mathcal{D}_T$:
  - $\mathcal{X}_S \neq \mathcal{X}_T$: **Heterogenous TL**
  - $\mathcal{X}_S = \mathcal{X}_T$ **but** $P(X_S) \neq P(X_T)$: **Homogenous TL**
- If $\mathcal{T}_S \neq \mathcal{T}_T$:
  - $\mathcal{Y}_S \neq \mathcal{Y}_T$: **Mismatch in label space**
  - $\mathcal{Y}_S = \mathcal{Y}_T$ **but** $f_S(.) \neq f_T(.)$ **Different predictive function** ($P(Y_S|X_S) \neq P(Y_T|X_T)$ conditional probability distributions)

Domain differences vs task differences [Pan et al. 2010]

Domain and Task differences between Target and Source.

Learning Settings		Source and Target Domains	Source and Target Tasks
Traditional Machine Learning		the same	the same
Transfer Learning	*Inductive Transfer Learning* /	the same	different but related
	Unsupervised Transfer Learning	different but related	different but related
	Transductive Transfer Learning	different but related	the same

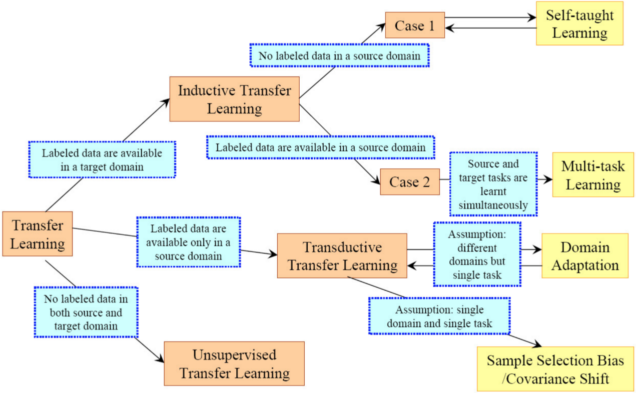
Corresponding **Labels Availability**.

Transfer Learning Settings	Related Areas	Source Domain Labels	Target Domain Labels	Tasks
Inductive Transfer Learning	Multi-task Learning	Available	Available	Regression, Classification
	Self-taught Learning	Unavailable	Available	Regression, Classification
Transductive Transfer Learning	Domain Adaptation, Sample Selection Bias, Co-variate Shift	Available	Unavailable	Regression, Classification
Unsupervised Transfer Learning		Unavailable	Unavailable	Clustering, Dimensionality Reduction

Domain differences vs task differences [Pan et al. 2010]

Relation with other techniques

Learning Settings		Source and Target Domains	Source and Target Tasks
Traditional Machine Learning		the same	the same
Transfer Learning	*Inductive Transfer Learning /*	the same	different but related
	Unsupervised Transfer Learning	different but related	different but related
	Transductive Transfer Learning	different but related	the same



Outline

Feature space differences

"Process of **improving the target predictive function** $f_T(.)$ by using the **related information** from $\mathcal{D}_S$ and $\mathcal{T}_S$" [Weiss et al. 2016]

Where: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$

- If $\mathcal{D}_S \neq \mathcal{D}_T$:
  - $\mathcal{X}_S \neq \mathcal{X}_T$: **Heterogenous TL**
  - $\mathcal{X}_S = \mathcal{X}_T$ **but** $P(X_S) \neq P(X_T)$: **Homogenous TL**
- If $\mathcal{T}_S \neq \mathcal{T}_T$:
  - $\mathcal{Y}_S \neq \mathcal{Y}_T$: **Mismatch in label space**
  - $\mathcal{Y}_S = \mathcal{Y}_T$ **but** $f_S(.) \neq f_T(.)$ **Different predictive function** ($P(Y_S|X_S) \neq P(Y_T|X_T)$ conditional probability distributions)

Homogeneous Transfer Learning: $\mathcal{X}_S = \mathcal{X}_T$:

- Correct $P(X_S) \neq P(X_T)$
- Correct $P(Y_S|X_S) \neq P(Y_T|X_T)$
- Correct both problems

Heterogeneous Transfer Learning: $\mathcal{X}_S \neq \mathcal{X}_T$:

- Align source and target feature spaces
- Correct probability distributions (homogeneous TL)

Outline

Methodology (what is being transferred)

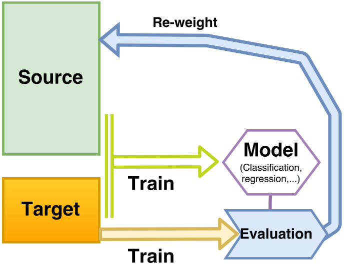
Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Reweight instances from X_S (e.g., [Jiang and Zhai 2007])

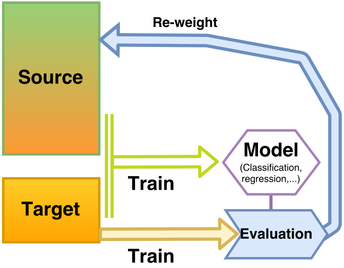


- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Reweight instances from X_S (e.g., [Jiang and Zhai 2007])

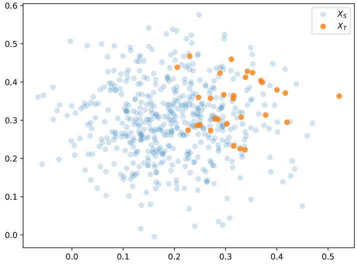


- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Reweight instances from X_S (e.g., [Jiang and Zhai 2007])

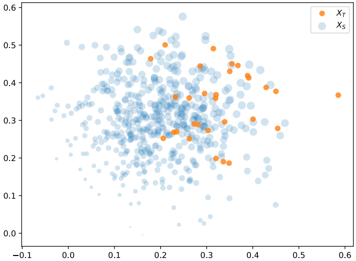


- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Reweight instances from X_S (e.g., [Jiang and Zhai 2007])

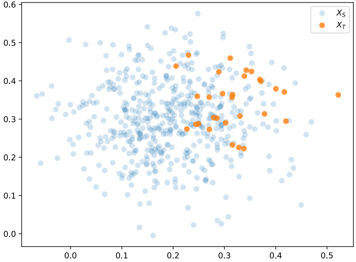


- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Sample instances from X_T

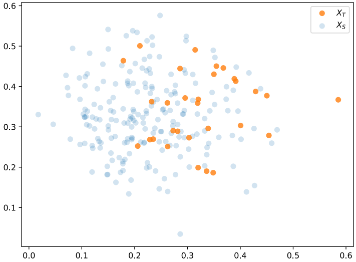


- **Feature-based**
- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based** Sample instances from X_T



- **Feature-based**
- **Parameter-based**
- **Relation-based**

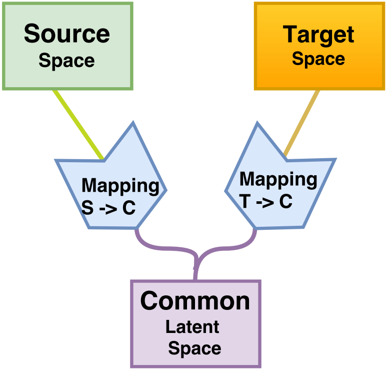
Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**

  Symmetric:

  Common latent feature space with predictive qualities.

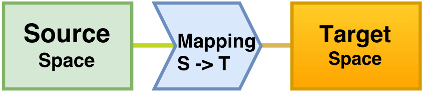


- **Parameter-based**
- **Relation-based**

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**

  Asymmetric: Map the source feature space to match target.



- **Parameter-based**
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**

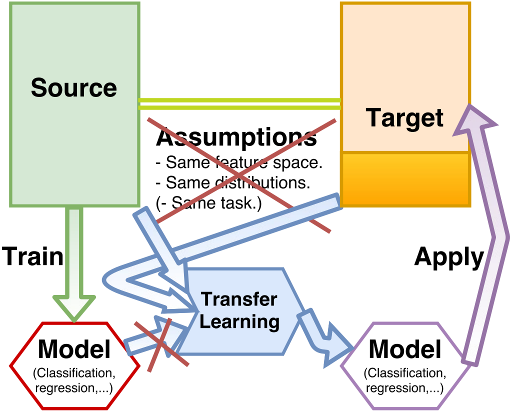
  Heterogeneous Symmetric example [Li et al. 2014]
  Semi-supervised heterogeneous feature augmentation (SHFA)

  - Instances: $x^S \in \mathbb{R}^{d_s}$, $x^T \in \mathbb{R}^{d_T}$, Labels: $y_i^S, y_i^T \in \{-1, 1\}$
  - Projection Matrices: $P \in \mathbb{R}^{d_c \times d_S}$, $Q \in \mathbb{R}^{d_c \times d_T}$
  - Mapping functions: $\phi_S(x^S) = \begin{bmatrix} Px^S \\ x^S \\ 0_{d_T} \end{bmatrix}$ and $\phi_T(x^T) = \begin{bmatrix} Qx^T \\ 0_{d_S} \\ x^T \end{bmatrix}$
  - Weight vector $w = [w_c, w_s, w_t]$
  - Search P, Q, w to optimize a hinge loss function (computed on source and target)
  - Based on Kernels, in practice.

- **Parameter-based**
- **Relation-based**

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**

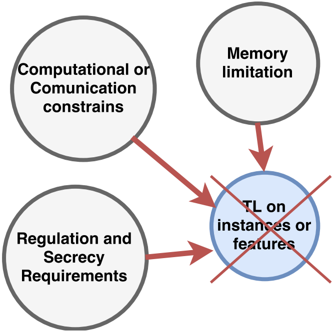


- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**

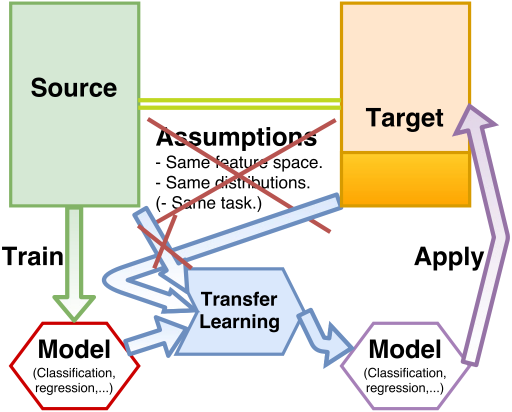


- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**



- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**
  - Bayesian algorithms (priors, ...).
  - Pre-trained deep-learning models (very trendy).
  - Other techniques: fewer (SVM, Random Forests).
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**
  TL for Random Forests [Segev et al. 2017]
  - Two kinds of changes are addressed:
    - Translations or shifts (method called STRUT)
    - Decision boundaries need to be refined/coarse-grained (method called SER)
  - Assumptions: $\mathcal{X}_S = \mathcal{X}_T$ and $\mathcal{Y}_S = \mathcal{Y}_T$
  - Create one forest with STRUT, one with SER $\rightarrow$ MIX forest
- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

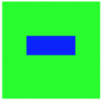
- **Instance-based**
- **Feature-based**
- **Parameter-based**
  TL on RF: Structure Expansion Reduction (SER):
  - Expansion: Expand leaf using the target instances.
  - Reduction: Reduce internal nodes if error decreases.
  - Expansion $\rightarrow$ reduction (intuition: keep new model similar to original)
- **Relation-based**

Methodology (what is being transferred)

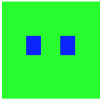
Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**

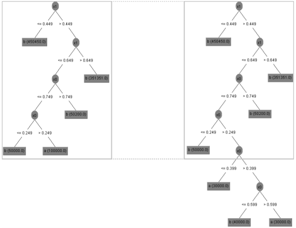
  TL on RF: Structure Expansion Reduction (SER):



(a) single box prior to splitting



(b) boxes after splitting



- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

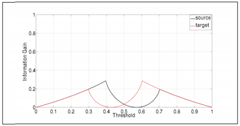
- **Instance-based**
- **Feature-based**
- **Parameter-based**
  TL on RF: Structure Transfer Algorithm (STRUT):
  - Discard thresholds.
  - Use Information Gain and D_{KL} w.r.t. old label distributions to select new thresholds.
  - Intuition: Decision Trees for similar problems tend to have similar structures
- **Relation-based**

Methodology (what is being transferred)

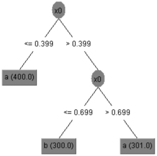
Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**

  TL on RF: Structure Transfer Algorithm (STRUT):



(a) IG value for thresholds



- **Relation-based**

Methodology (what is being transferred)

Four general Transfer Categories [Pan et al. 2010]

- **Instance-based**
- **Feature-based**
- **Parameter-based**
- **Relation-based**
    - Few approaches ([Mihalkova et al. 2007])
    - Designed for relational data.
    - Markov Logic Networks.

Director(jack) Actor(jill)
MovieMember(movie1, jack) MovieMember(movie1, jill)
MovieMember(movie2, jill) WorkedFor(jill, jack)

Figure 2: Example relational database

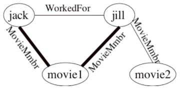


Figure 3: Example of a relational graph

Outline

Active research topic with many related tasks

- Many related terms [Pan et al. 2010]:
    - Learning to learn
    - Life-long learning
    - Incremental/online/cumulative learning
    - Knowledge transfer
    - Multi-task learning
    - Domain adaptation
    - Inductive transfer
    - Transductive transfer
    - Context-sensitive learning
    - Meta learning
    - (Pre-trained models)
    - ...
- **700** new papers between 2010 and 2016 [Weiss et al. 2016] .

Outline

1. Introduction
   - General concept
   - Transfer Learning in Machine Learning
   - Formal definition
2. Transfer Learning Taxonomy
   - Labels availability
   - Domain differences vs task differences
   - Feature space differences
   - Methodology (what is being transferred)
3. Related tasks
   - Self-taught learning
   - Multi-task learning
   - Adaptation of probability distributions
4. Negative Transfer

- Use **large unlabelled dataset** $\{x_u^{(1)}, ..., x_u^{(k)}\}$ (s.t., $x_u^{(i)} \in \mathbb{R}^n$) to build **basis vectors** $b = \{b_1, b_2, ..., b_s\}$ by solving:

$$minimize_{a,b} \sum_i ||x_u^{(i)} - \sum_j a_j^{(i)} b_j||_2^2 + \beta ||a^{(i)}||_1$$

With $||b_j||_2 \leq 1$ and $a^{(i)} \in \mathbb{R}^s$: i-th instance **basis activation**.

- Compute **features** for **train** set $\{(x_l^{(1)}, y^{(1)}), ..., (x_l^{(m)}, y^{(m)})\}$:

$$\hat{a}(x_l^{(i)}) = argmin_{a^{(i)}} ||x_l^{(i)} - \sum_j a_j^{(i)} b_j||$$

- **Train an algorithm** using the $\hat{a}(x_l^{(i)})$ as **features**.
- Images, Audio signals, Text...

Outline

- "Training **tasks in parallel** while using a **shared representation**" [Caruana 1997].
- **Deep Learning Approaches**:
  - Hard parameter sharing: First layers shared (e.g., [Baxter 1997]).
  - Soft parameter sharing: Each model has its own parameters but regularisation ensures similarity (e.g., [Duong et al. 2015]).
  - More recent approaches: Combine Hard / soft / task relationship (e.g., [Long et al. 2015] [Lu, 2016] [Misra, 2016]).
- Learn **tasks relationship** (e.g., cluster the tasks [Jacob et al. 2009])
- **Reproducing Hilbert spaces** (e.g., [Ciliberto et al. 2015]).

Assumption Predictive functions from different domains share a small set of features

- $f_t(x) = \sum_{i=1}^{d} a_{i,t} h_i(x)$, with $h_i(x) = \langle u_i, x \rangle$
- Matrix U is a $d \times d$ orthogonal with u_i as i-th column.

Objective Function

$$\sum_{t \in \{T,S\}} \sum_{i=1}^{n_t} L(y_{t,i}, \langle a_t, U^\top x_{t,i} \rangle) + \gamma ||a_t||_1^2$$

And $a_T U^\top X_T$ and $a_S U^\top X_S$: low dimensional representations

Outline

- **Domain adaptation** [Daumé and Marcu 2006], **sample selection bias** [Zadrozny 2004] **co-variate shift** [Shimodaira 2000].
- "**Alter** the **source** domain to bring the **distribution** of the source **closer** to that of the **target**" [Weiss et al. 2016]
- TL Case were $\mathcal{X}_S = \mathcal{X}_T$, $P(X_S) \neq P(X_T)$ and same predictive function.
- Goal: Estimate $\frac{P(X_S)}{P(X_T)}$ (kernel methods, D_{kl}, ...)
- Well established task. [book by Quionero-Candela et al. 2009]

Outline

Negative TL

"Information learned from a **source** domain has **detrimental** effect on the **target learner**" [Weiss et al. 2016]

- $f_{T1}(.)$ trained only with $\mathcal{D}_T$.
- $f_{T2}(.)$ trained with $\mathcal{D}_T$ and $\mathcal{D}_S$.
- **Negative Transfer**:
  $Performance(f_{T1}(.)) > Performance(f_{T2}(.))$
- Negative Transfer decreases when the number of labeled target samples increases [Rosentein et al. 2005]
- Measure "transferability" between source and target, cluster similar domains, use only most transferable domains (e.g., [Ge et al. 2013][Eaton et al. 2008])
- Few works, and interesting topic.