迁移学习及其应用
Introduction to Transfer Learning

杨强，Qiang Yang

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Hong Kong

http://www.cse.ust.hk/~qyang

Transfer Learning? (DARPA 05)

Herb Simon defined learning as:

"Any change in a system that allows it to perform better the second time on repetition of the same task or on another task ***drawn from the same distribution.****"* (1983)

- This has been the predominant task of machine learning research
- In contrast, people often transfer knowledge to novel situations
  - Chess → checkers
  - C++ → Java
  - Physics → Computer Science

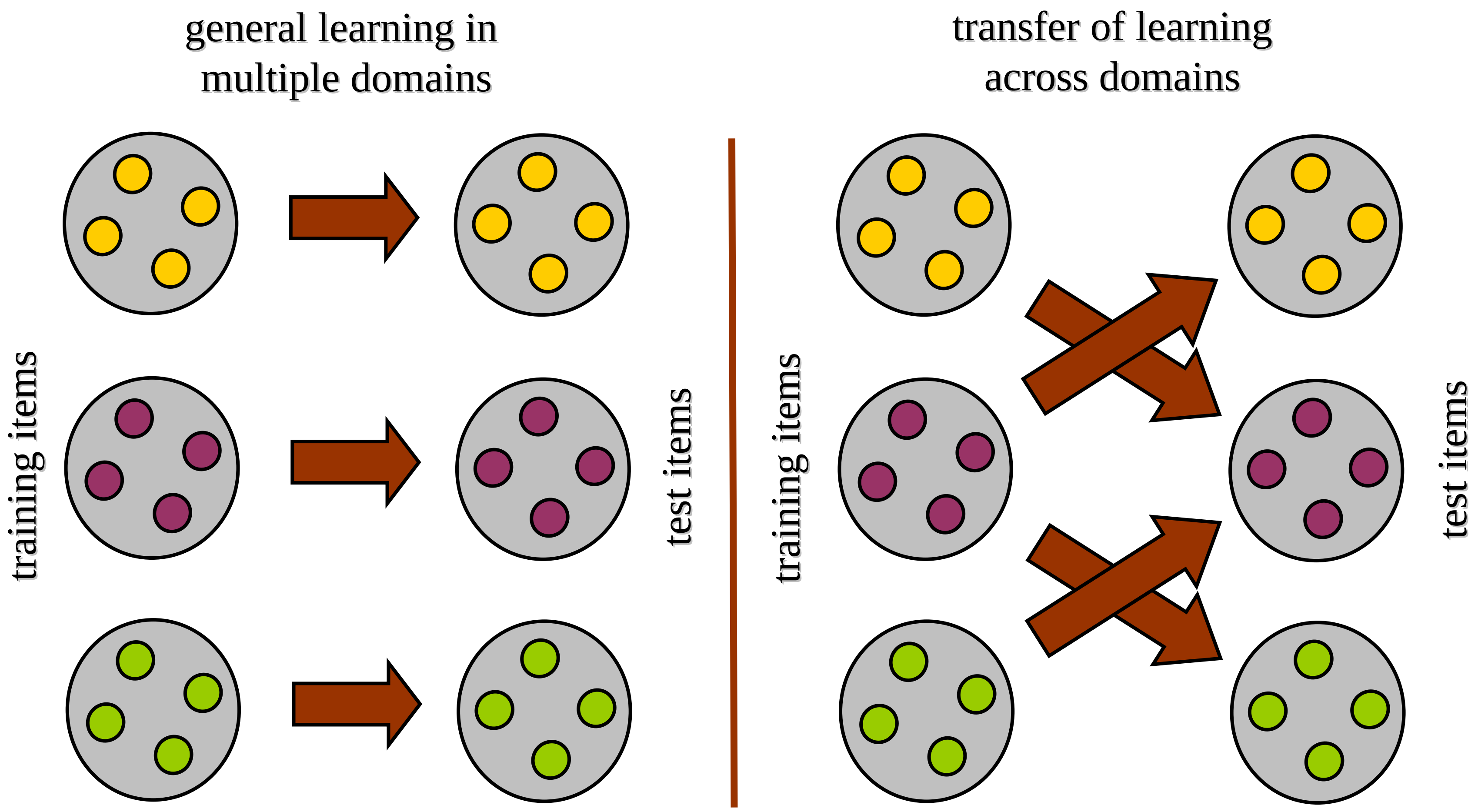
Transfer Learning:

The ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks (or new domains)

Machine Learning...

- Traditional machine learning
  - 种瓜得瓜，种豆得豆
- Transfer Learning 迁移学习
  - 举一反三
  - 投桃报李

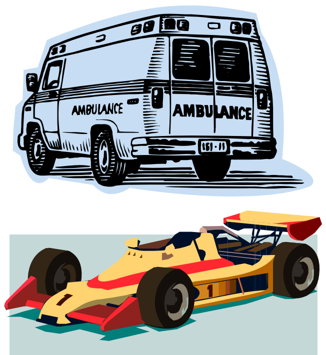
Generality and Transfer In Learning
(P. Langley 06)



Humans can learn in many domains.

Humans can also transfer from one domain to other domains.

Domain Classes That Exhibit Transfer (Langley 06)



From: tsenator@darpa.mil
To: langley@csli.stanford.edu
Subject: site visit next week
Date: Nov 14, 2004

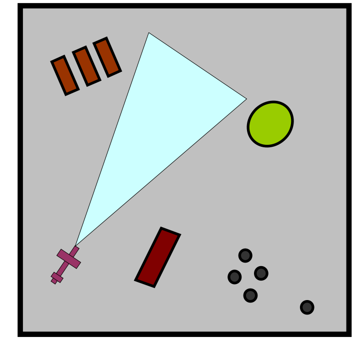
Pat – I am looking forward to hearing about your progress over the past year during my site visit next week. - Ted

From: noname@somewhere.com
To: langley@csli.stanford.edu
Subject: special offer!!!
Date: Nov 14, 2004

One week only! Buy v*i*a*g*r*a at half the price available in stores. Go now to http://special.deals.com

Which is an emergency vehicle? Which email is spam?

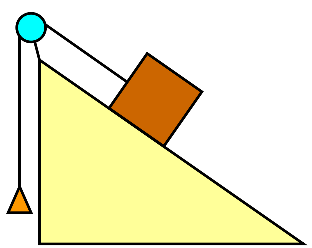
654	456	821
- 321	- 237	- 549
940	601	400
- 738	- 459	- 321



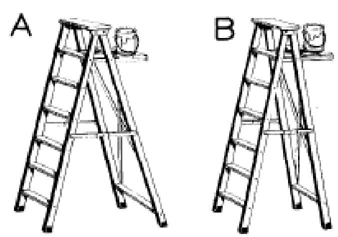
What are the problem answers? What path should the plane take?

Classification tasks that involve assigning items to categories, such as recognizing types of vehicles or detecting spam.

Procedural tasks that involve execution of routinized skills, both cognitive (e.g., multi-column arithmetic) and sensori-motor (e.g., flying an aircraft).

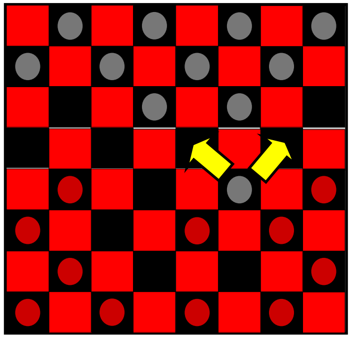


A block sits on an inclined plane but is connected to a weight by a string through a pulley. If the angle of the plane is 30 degrees and . . .



Which ladder is safer to climb on?

What should the blue team do?



Which jump should red make?

Inference tasks that require multi-step reasoning to obtain an answer, such as solving physics word problems and aptitude/achievement tests.

Problem-solving tasks that benefit from strategic choices and heuristic search, such as complex strategy games.

Why Transfer Learning?

- Nature is like that
  - training and testing data often have different distributions
- Economics
  - We have large amounts of labeled data or trained classifiers
    - Why waste old data?
    - Re-use old labelled data to save costs
- Efficiency
  - Wish to learn faster

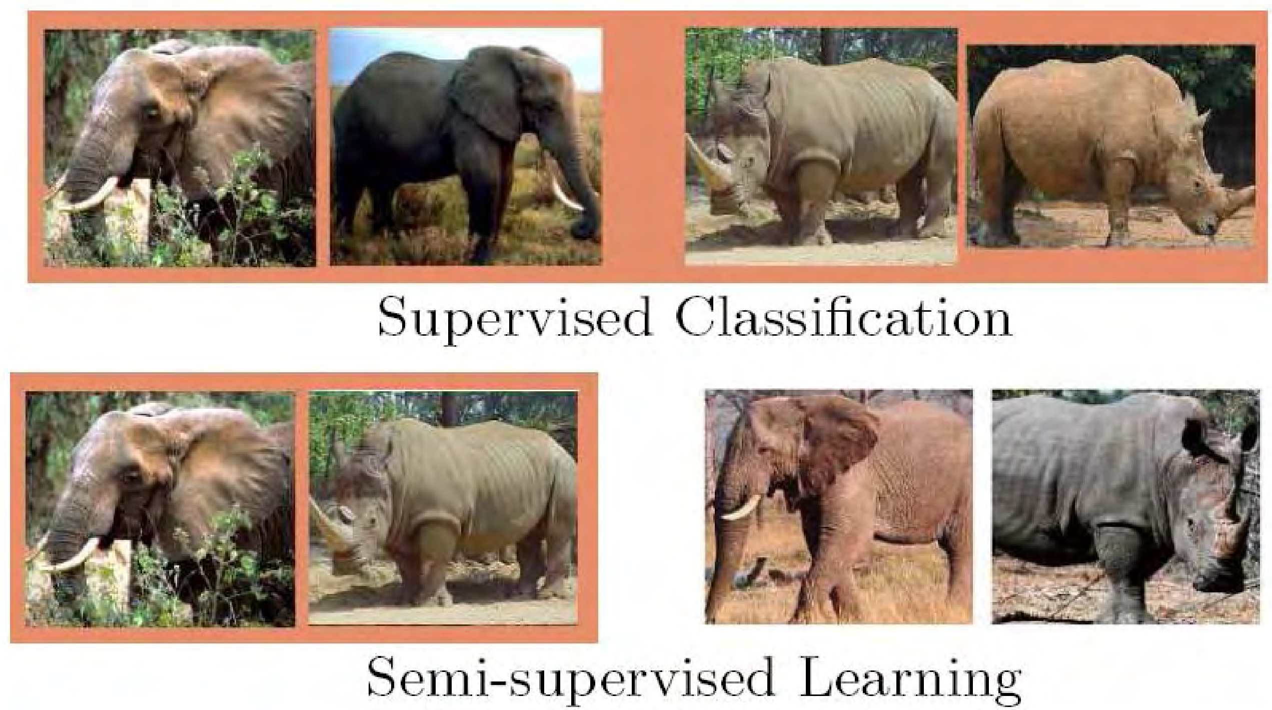
Progress Toward Reducing Learning Efforts

Supervised Classification

Transfer Learning

(from Raina et al. ICML 06)

Progress Toward Reducing Learning Efforts



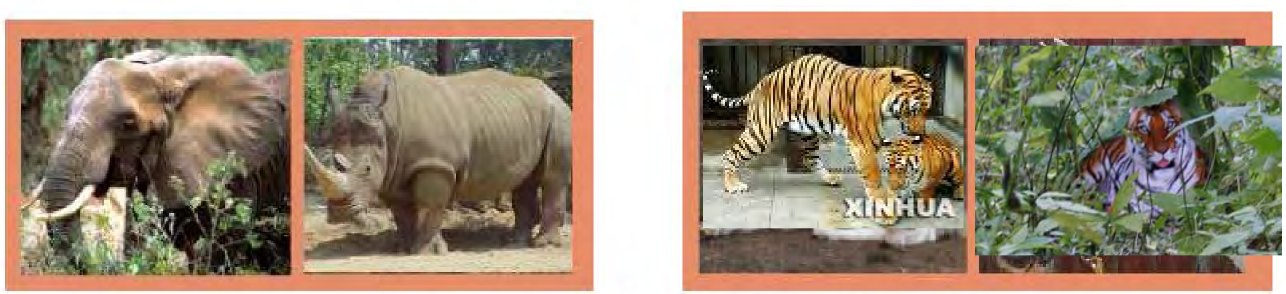
Transfer Learning

(from Raina et al. ICML 06)

Progress Toward Reducing Learning Efforts

Supervised Classification

Semi-supervised Learning



Transfer Learning

(from Raina et al. ICML 06)

Types of Transfer Learning

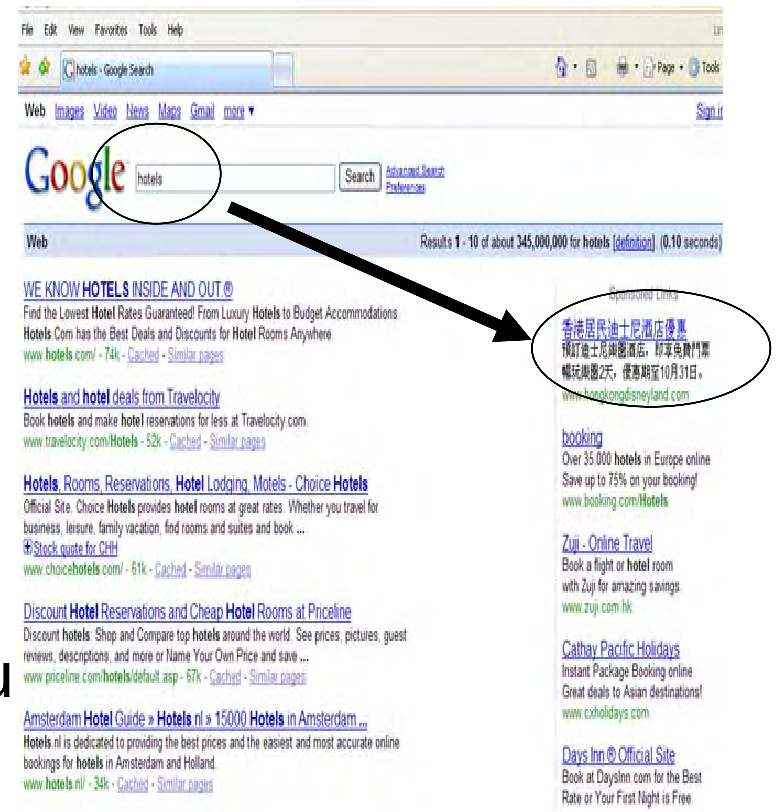
- source $\neq$ target
  - $Pr_s(X) \neq Pr_t(X)$:
    - sample selection bias (Zadrony04)
  - $Pr_s(Y) \neq Pr_t(Y)$:
    - class imbalance problem (Elkan00)
  - $Pr_s(Y/X) \neq Pr_s(Y/X)$:
    - concept drift (Widmer96)
  - $Pr_s(X,Y) \neq Pr_t(X,Y)$:
    - domain transfer learning

Case 1: 目标变化→ 目标迁移

- Target Class Changes → Target Transfer Learning
  - Solution: 以不变应万变

Query Classification and Online Advertisement

- ACM KDDCUP 05 Winner
- SIGIR 06
- ACM Transactions on Information Systems Journal 2006
  - Joint work with Dou Shen, Jiantao Sun and Zheng Chen



QC as Machine Learning

Ins ... tition

... gories

... gines

... with

Table I. Examples of queries.
1967 shelby mustang
actress hildegarde
Aldactone
alfred Hitchcock
amazon rainforest
section8rentalhouses.com
Sakpsabancnnhayat

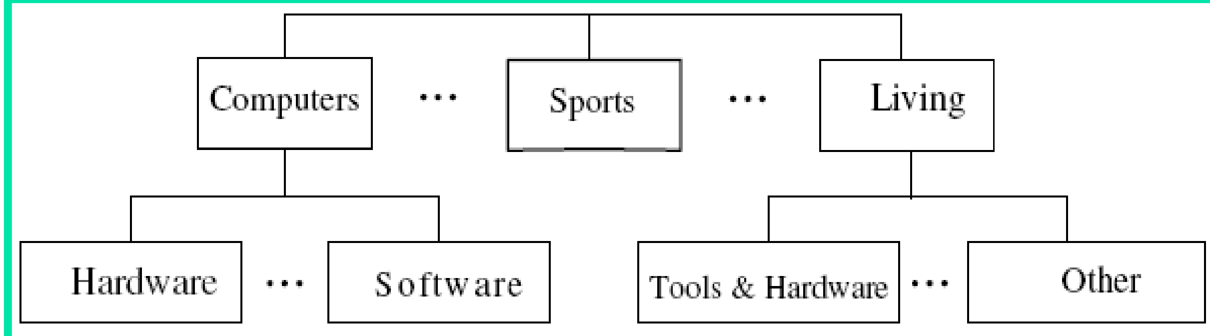


Figure 1: An Example of the Target Taxonomy.

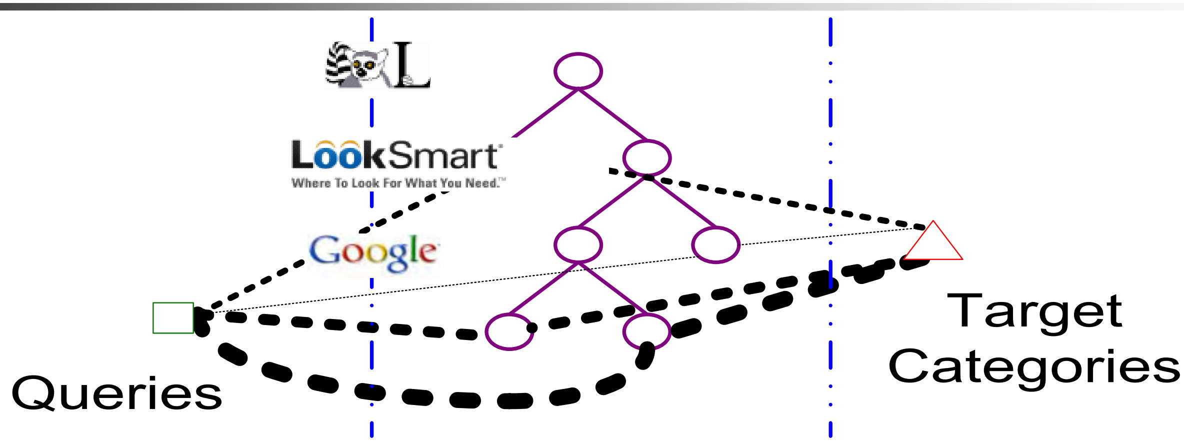
Related Works

- Document/Query Expansion
  - Borrow text from extra data source
    - Using hyperlink [Glover 2002];
    - Using implicit links from query log [Shen 2006];
    - Using existing taxonomies [Gabrilovich 2005];
  - Query expansion [Manning 2007]
    - Global methods: independent of the queries
    - Local methods using relevance feedback or pseudo-relevance feedback
- Query Classification/Clustering
  - Classify the Web queries by geographical locality [Gravano 2003];
  - Classify queries according to their functional types [Kang 2003];
  - Beitzel et al. studied the topical classification as we do. However they have manually classified data [Beitzel 2005];
  - Beeferman and Wen worked on query clustering using clickthrough data respectively [Beeferman 2000; Wen 2001];

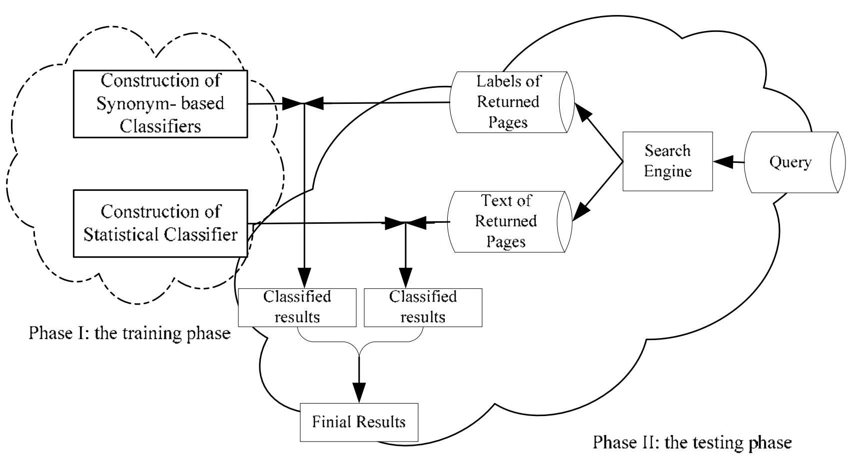
Target-transfer Learning in QC

- Classifier, once trained, stays constant
  - Target Classes Before
    - Sports, Politics (European, US, China)
  - Target Classes Now
    - Sports (Olympics, Football, NBA), Stock Market (Asian, Dow, Nasdaq), History (Chinese, World) How to allow target to change?
- Application:
  - advertisements come and go,
  - but our query→target mapping needs not be retrained!
- We call this the target-transfer learning problem

Solutions: Query Enrichment + Staged Classification

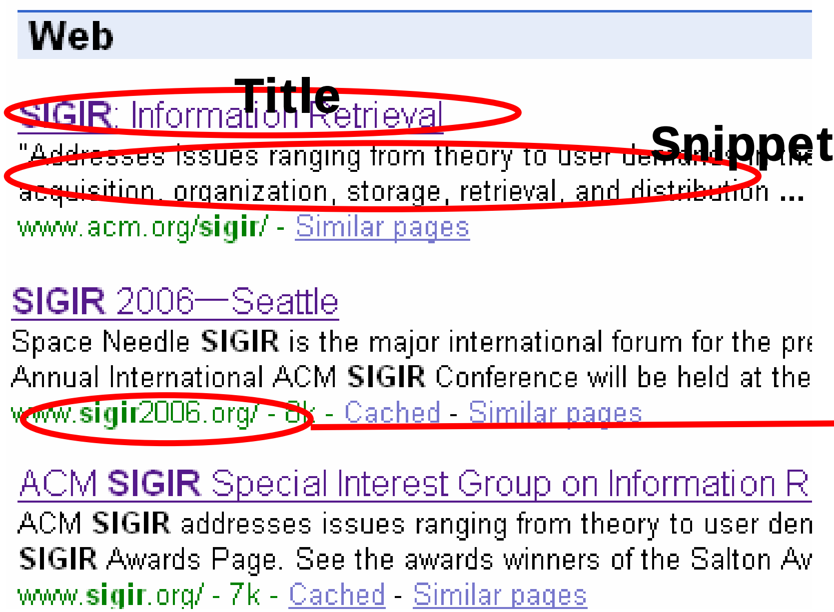


Solution: Bridging classifier



Step 1: Query enrichment

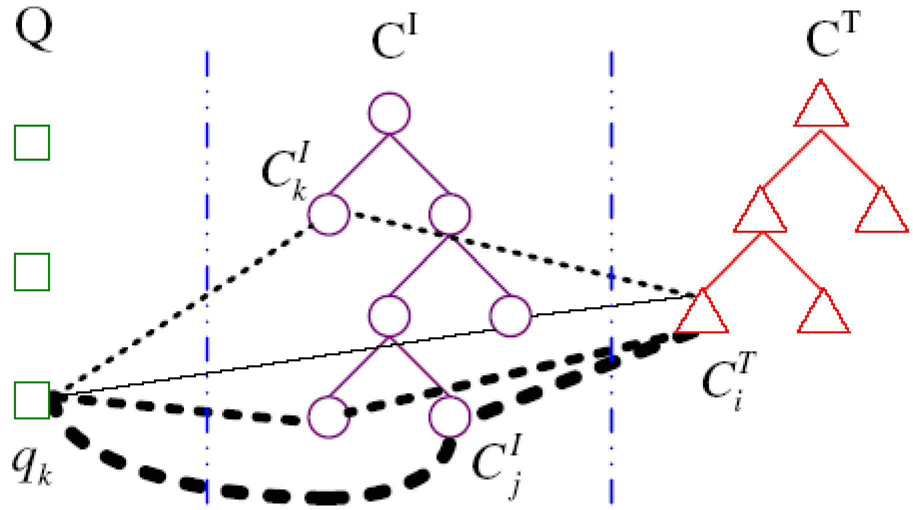
- Textual information
- Category information



Full text

Step 2: Bridging Classifier

- Wish to avoid:
  - When target is changed, training needs to repeat!
- Solution:
  - Connect the target taxonomy and queries by taking an intermediate taxonomy as a bridge



Bridging Classifier (Cont.)

- How to connect?

$$
\begin{aligned}
p(C_i^T|q) &= \sum_{C_j^I} p(C_i^T, C_j^I|q) \\
&= \sum_{C_j^I} p(C_i^T|C_j^I, q)p(C_j^I|q) \\
&\approx \sum_{C_j^I} p(C_i^T|C_j^I)p(C_j^I|q) \\
&= \sum_{C_j^I} p(C_i^T|C_j^I)\frac{p(q|C_j^I)p(C_j^I)}{p(q)} \\
&\propto \sum_{C_j^I} p(C_i^T|C_j^I)p(q|C_j^I)p(C_j^I)
\end{aligned}
$$

The relation between C_i^T and C_j^I

The relation between q and C_j^I

Prior prob. of C_j^I

The relation between q and C_i^T

$$\boxed{c^* = \arg\max_{C_i^T} p(C_i^T|q)}$$

Category Selection for Intermediate Taxonomy

- Category Selection for Reducing Complexity
  - Total Probability (TP)

$$Score(C_j^I) = \sum_{C_i^T} \hat{P}(C_i^T | C_j^I)$$

  - Mutual Information

$$MI(C_i^T, C_j^I) = \frac{1}{|C_i^T|} \sum_{t \in C_i^T} MI(t, C_j^I)$$

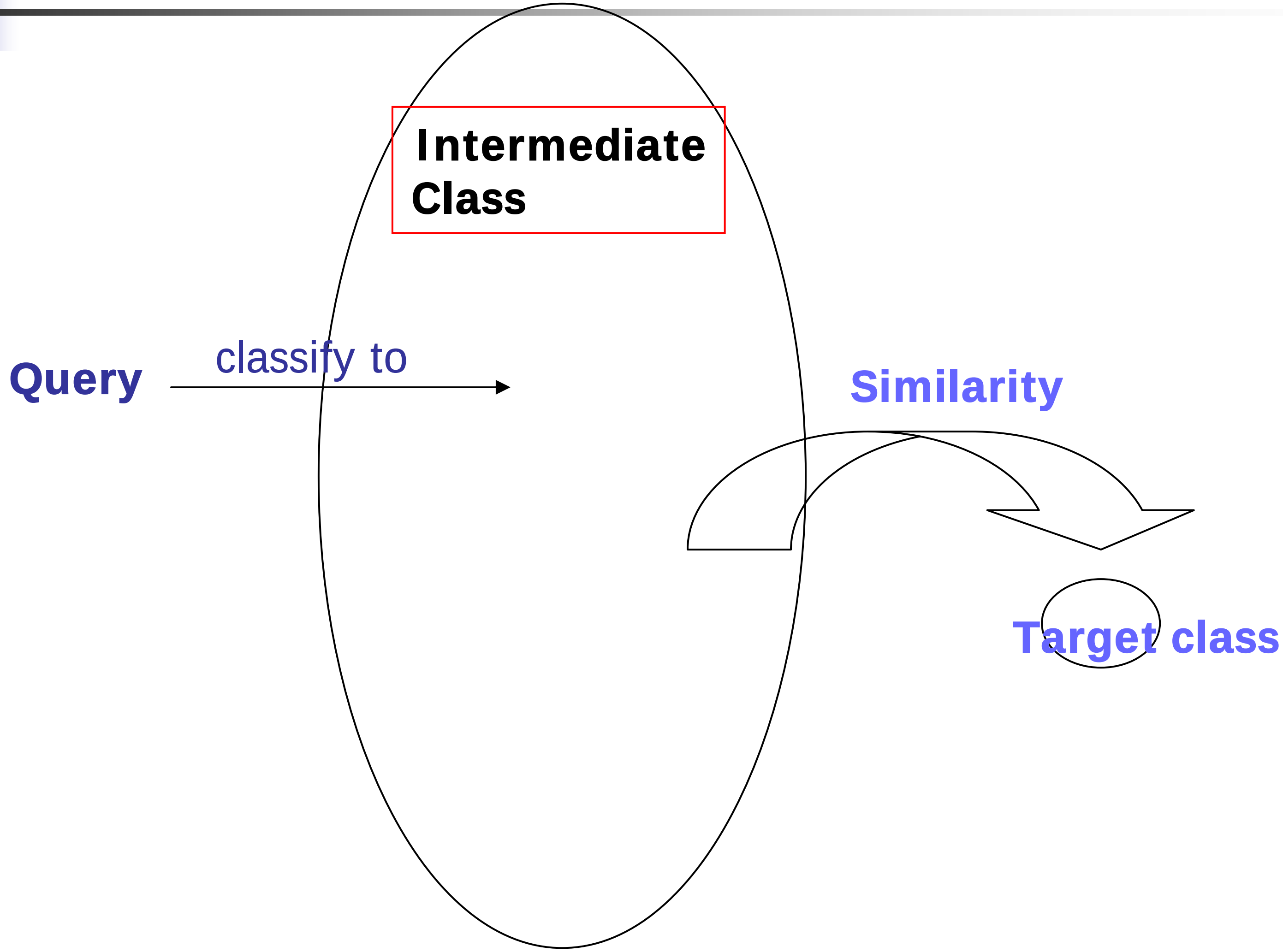
$$MI_{avg}(C_j^I) = \sum_{C_j^T} MI(C_i^T, C_j^I)$$

Experiment — Data Sets & Evaluation

- ACM KDDCUP
  - Starting 1997, ACM KDDCup is the leading Data Mining and Knowledge Discovery competition in the world, organized by ACM SIG-KDD.
- ACM KDDCUP 2005
  - Task: Categorize 800K search queries into 67 categories
  - Three Awards
  - (1) Performance Award ; (2) Precision Award; (3) Creativity Award
  - Participation
    - 142 registered groups;
    - 37 solutions submitted from 32 teams
- Evaluation data
  - 800 queries randomly selected from the 800K query set
  - 3 human labelers labeled the entire evaluation query set
- Evaluation measurements: Precision and Performance (F1)
- <u>We won all three</u>.

$$\text{Overall F1} = \frac{1}{3}\sum_{i=1}^{3}(\text{F1 against human labeler i})$$

Summary: Target-Transfer Learning



Case 2: Domain Transfer Learning

- Q: “What if the source and target domain distributions are different?”
  - Joint work with Arthur Dai, G. Xue and Yong Yu.
  - ACM KDD 2007, ICML 2007, AAAI 2007, etc.

Training and Target difference in the real world

- 20 newsgroups (20,000 documents, 20 data sets)

Old

comp.graphics (**comp**)
comp.os.mis-windows.misc (**comp**)
sci.crypt (**sci**)
sci.electronics (**sci**)

→

New

comp.sys.ibm (?)
comp.windows.x (?)
sci.med (?)
sci.space (?)

- SRAA (A. McCallum, 70,000 articles)

Old

sim-auto (**auto**)
sim-aviation (**aviation**)

→

real-auto (?)
real-aviation (?)

New

- Reuters-21578

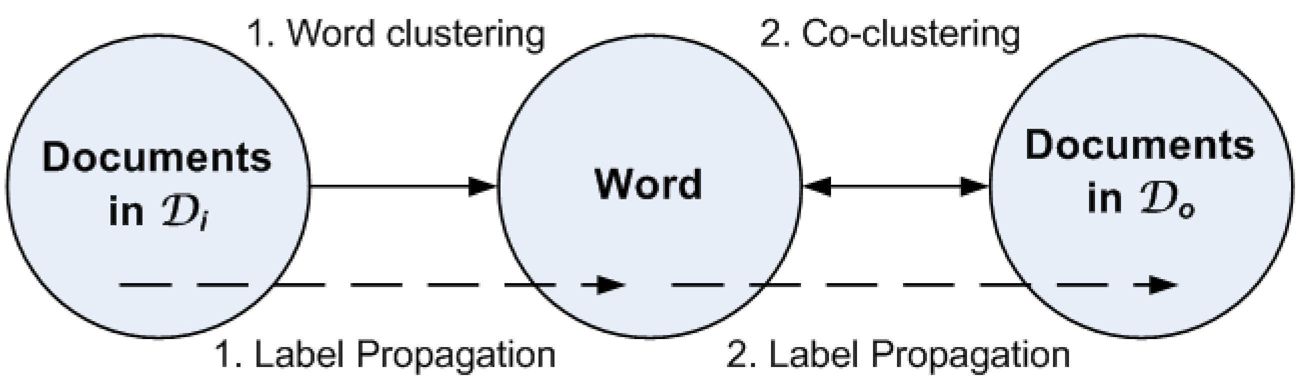
We have been working on this over the past year...

- ACM KDD '07 Presentation
- Feature Based Transfer Learning
  - Co-clustering based Classification
- Experimental Results in text mining

- Other works we've done
  - Instance Based Transfer Learning
  - Feature Based Transfer Learning
  - Embedded Transfer Learning
  - Semantic Structure Based Transfer Learning

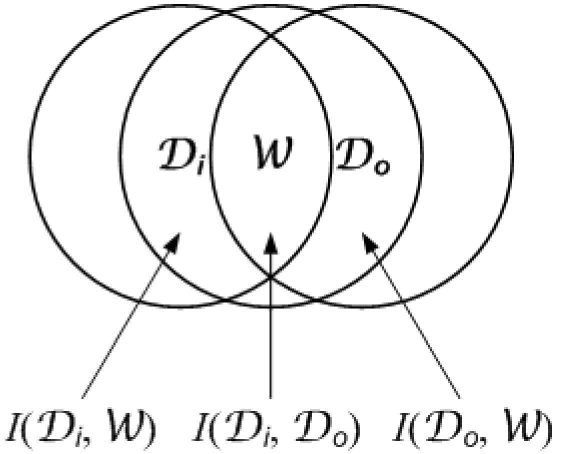
Feature-based Domain transfer

- Co-clustering is applied between features (words) and target-domain documents
- Word clustering is constrained by the labels of in-domain (Old) documents
  - The word clustering part in both domains serve as a ***bridge***



Label Propagation

- Co-clustering requires optimization
  - Objective function: based on mutual information *MI*(Partition 1, Partition 2)
  - When $I(D_i, W)$ and $I(D_o, W)$ are increasing, i.e. more dependent, $I(D_i, D_o)$ is likely to be non-decreasing, i.e., dependent



Optimization Function

- $\hat{\mathcal{D}}_o$ – clusters (classification) w.r.t. the target-domain documents
- $\hat{\mathcal{W}}$ – clusters w.r.t. the common features (words)
- C – class-labels of the source-domain documents
- Optimization Function

$$I(\mathcal{D}_o; \mathcal{W}) - I(\hat{\mathcal{D}}_o; \hat{\mathcal{W}}) + \lambda \cdot (I(\mathcal{C}; \mathcal{W}) - I(\mathcal{C}; \hat{\mathcal{W}}))$$

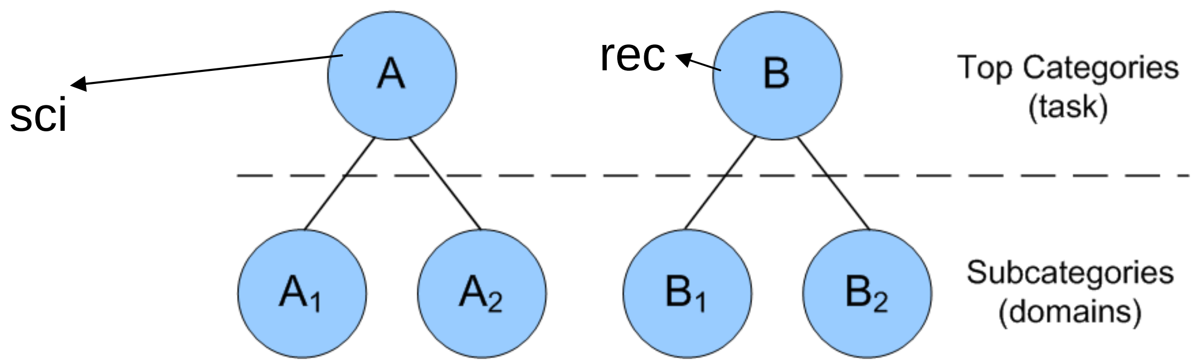
  - Minimize the loss in mutual information before and after clustering/classification
    - Between D_o and W
    - Between C and W

Ideas behind our Algorithm CoCC

- Co-Clustering-based Classification
- Iteratively choose the locally best doc-cluster/word-cluster
    1. cluster each document d to D_o and
    2. cluster each word w to W
- Objective:
    - reach the objective function through local optimization

Data Sets

- Three text collections
  - 20 newsgroups
  - SRAA
  - Reuters-21578
- The data are split based on sub-categories



Old Domain: [A1=+, B1= –],   New Domain: [A2=?, B2=?]

Document-word Co-occurrence

source

target

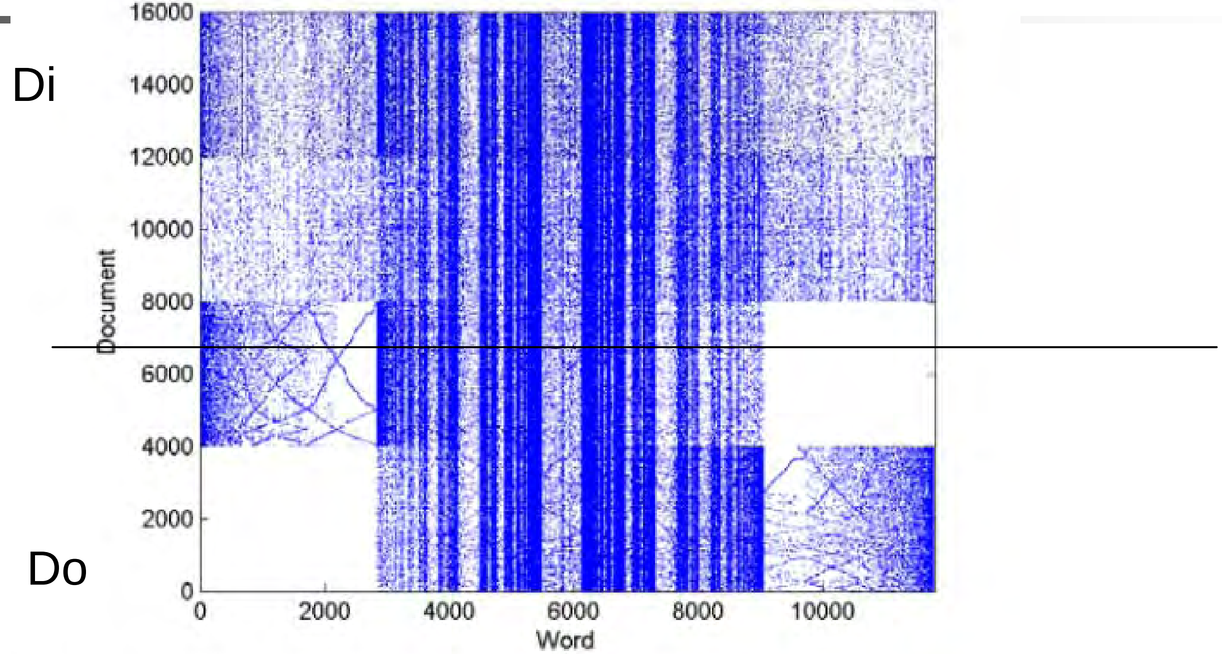


Figure 2: Document-word co-occurrence distribution on the auto vs aviation data set

Conclusions: D_i and D_o are similar but different

Performance

Transductive SVM (TSVM)
Spectral Graph Transducer (SGT)

- In test error rate

Data Set	NBC	SVM	TSVM	SGT	CoCC
real vs simulated	0.259	0.266	0.130	0.130	**0.120**
auto vs aviation	0.150	0.228	0.102	0.087	**0.068**
rec vs talk	0.235	0.233	0.040	0.091	**0.035**
rec vs sci	0.165	0.212	0.062	0.062	**0.055**
comp vs talk	0.024	0.103	0.097	0.028	**0.020**
comp vs sci	0.207	0.317	0.183	0.279	**0.130**
comp vs rec	0.072	0.165	0.098	0.047	**0.042**
sci vs talk	0.226	0.226	0.108	0.083	**0.054**
orgs vs places	0.377	0.454	0.436	0.385	**0.320**
people vs places	0.216	0.266	0.231	0.192	**0.174**
orgs vs people	0.289	0.297	0.297	0.306	**0.236**

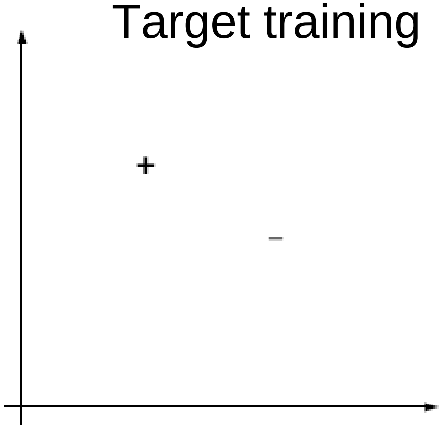
Conclusions: using CoCC can significantly reduce the error rates

Transferring Instances: Tradaboost

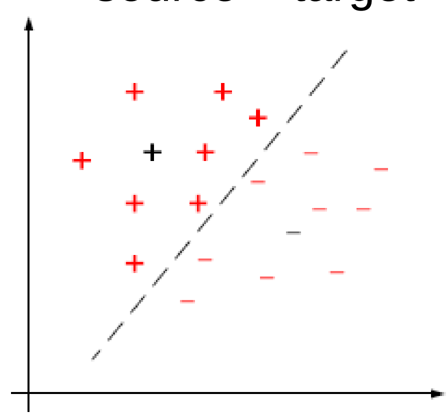
[Wu and Dietterich ICML 04] [Dai, Yang et al. ICML 07]

- Given
  - Insufficient labeled data from the target domain (primary data)
  - Labeled data following a different distribution (auxiliary data)
- The auxiliary data are weaker evidence for building the classifier

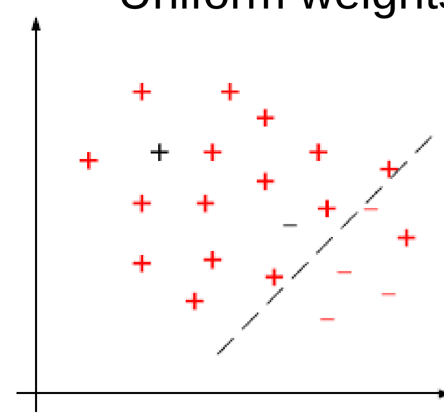
Target training



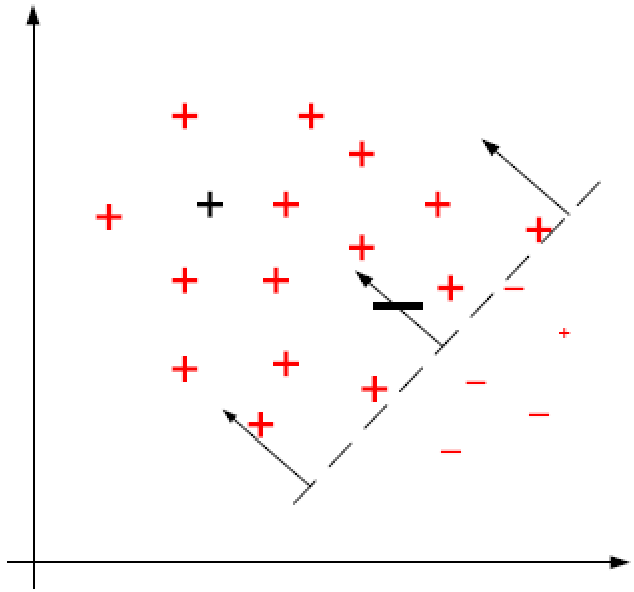
source + target



Uniform weights (X)



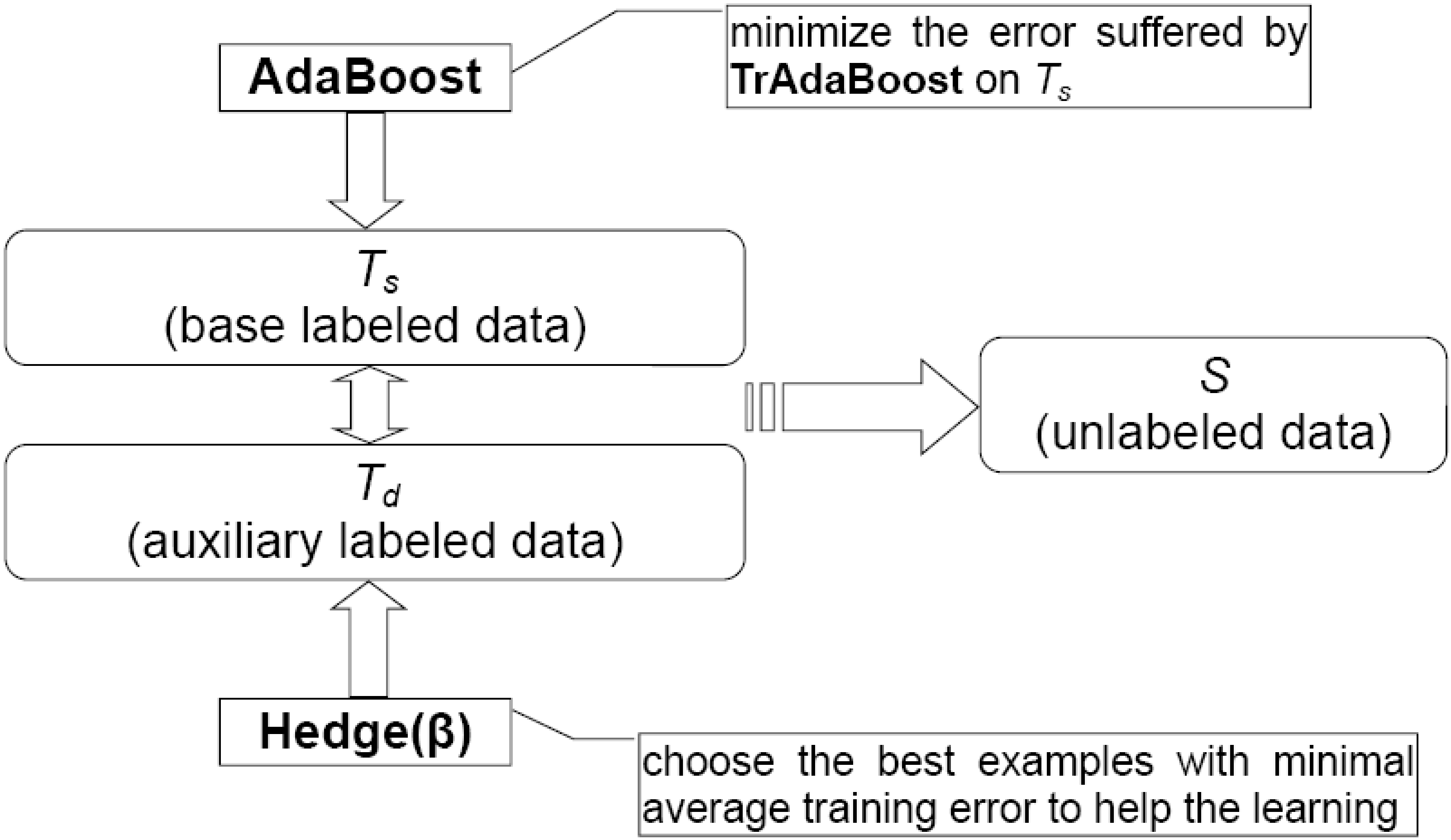
Incorporating Auxiliary (Source) Data into the Objective Function
(wu and dietterich 04, Dai et al. 07)



- Differentiate the cost for misclassification of

$$J'(h) = \sum_{i}^{N^p} L(h(\mathbf{x}_i^p), y_i^p) + \gamma \sum_{i}^{N^a} L(h(\mathbf{x}_i^a), y_i^a) + \lambda D(y)$$

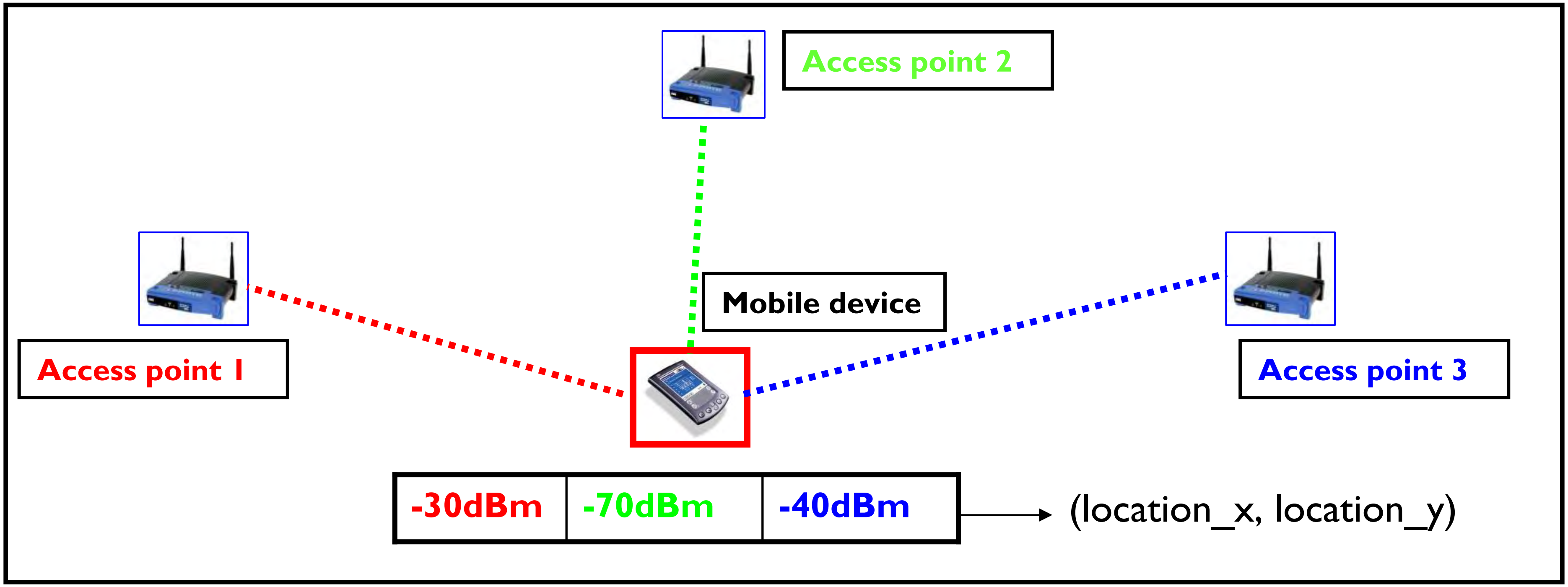
Tradaboost: The Main Idea



Latent Space –based Transfer Learning:

Localization in a WiFi Environment Through Transfer Learning [Pan, Yang et al. AAAI 07]

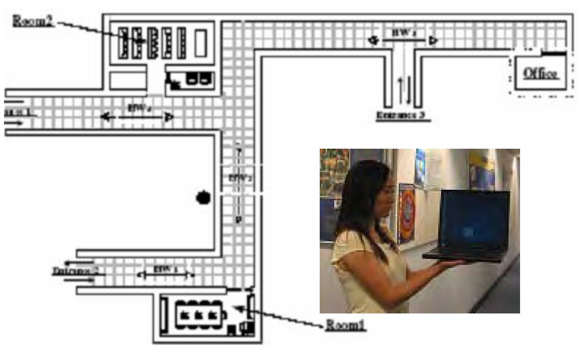
- Received-Signal-Strength (RSS) based localization in an Indoor WiFi environment.

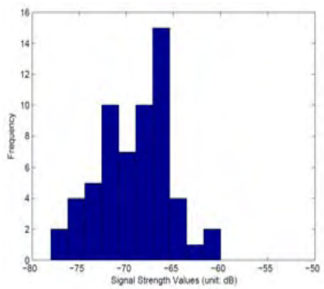


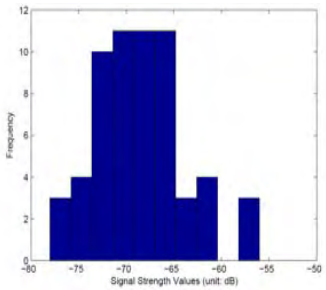
Where is the mobile device?

Distribution Changes

- The mapping function f learned in the offline phase can be out of date.
- **Recollecting the WiFi data is very expensive.**
- **How to adapt the model ?**







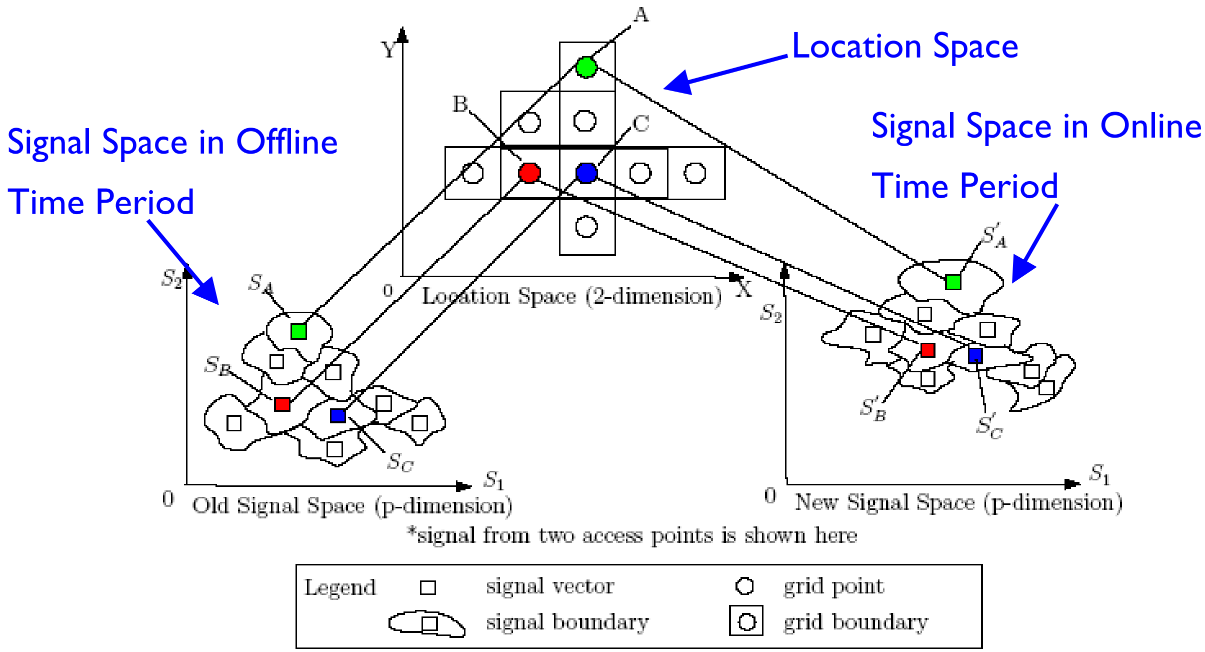
Time

Night time period t_0

Day time period t_1

Dynamic Localization Through Transfer Learning

- Since the location space does not change over time, different signal spaces have a common underlying low dimensional (2D/3D) manifold structure.



Reference points are placed at **A**, **B**, **C**, which bridge a connection between old signal space and new signal space.

Based on this connection, we can take into account the new signal data to adapt the old mapping function by transfer learning

How to solve the Transfer Learning problem?

We can learn a pair of functions $f = (f_{old}^*, f_{new}^*)$ together, such that:

$$f_{old}^*(S_{old}) = L$$

$$f_{new}^*(S_{new}) = L$$

$$A = f_{old}^*(S_A) = f_{new}^*(S_A')$$

$$B = f_{old}^*(S_B) = f_{new}^*(S_B')$$

$$C = f_{old}^*(S_C) = f_{new}^*(S_C')$$

Fitting a mapping function from old signal space to location space.

Fitting a mapping function from new signal space to location space.

The pair of functions should agree at corresponding pairs

Our Model-- LeManCoR

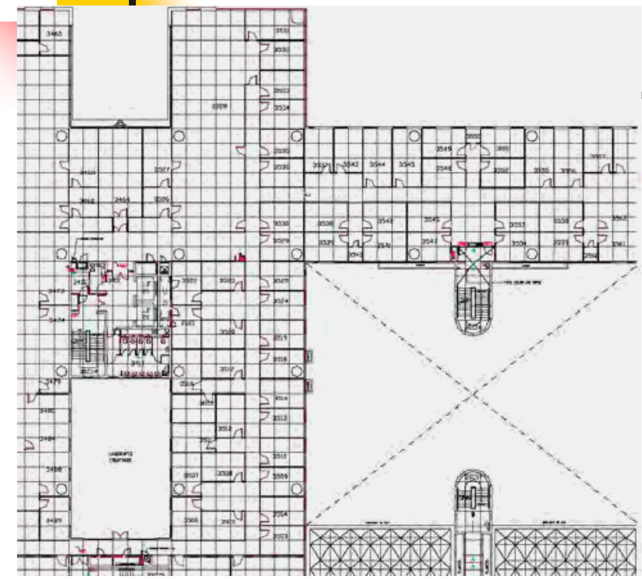
Adapted classifier function

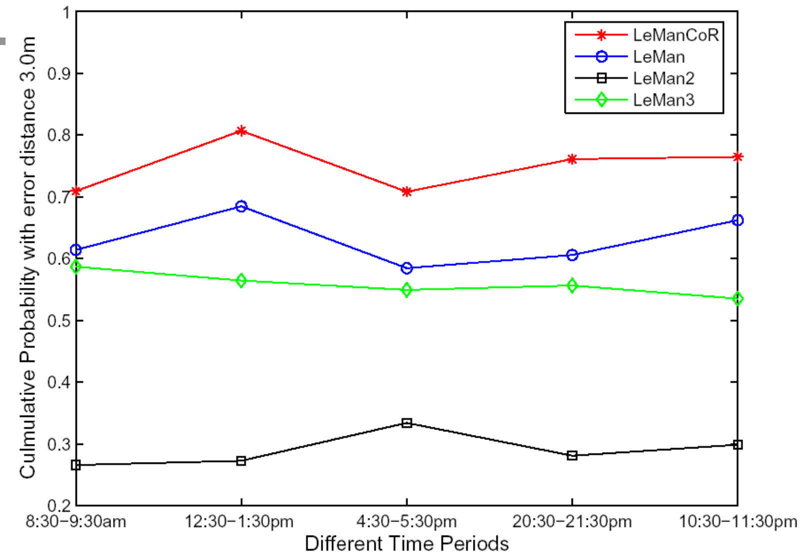
$$(f_{old}^*, f_{new}^*)$$

$$= \underset{f_{old}, f_{new}}{\operatorname{argmin}} \{ \| f_{old}(S_{old}) - L \| + \| f_{new}(S_{new}) - L \| + \sum_{r_i \in R} \| f_{old}(S_{r_i}) - f_{new}(S_{r_i}) \| \}$$

- The above equation is Manifold Regularization as the mapping function.
- This optimization problem is similar to manifold co-regularization (V. Sindhwani et al. 2005).
- The standard Manifold Co-Regularization approach cannot handle our case.
- We extend Manifold Co-Regularization to a more general case. It's localization version is called ***LeManCoR***.

Experimental Setup and Results

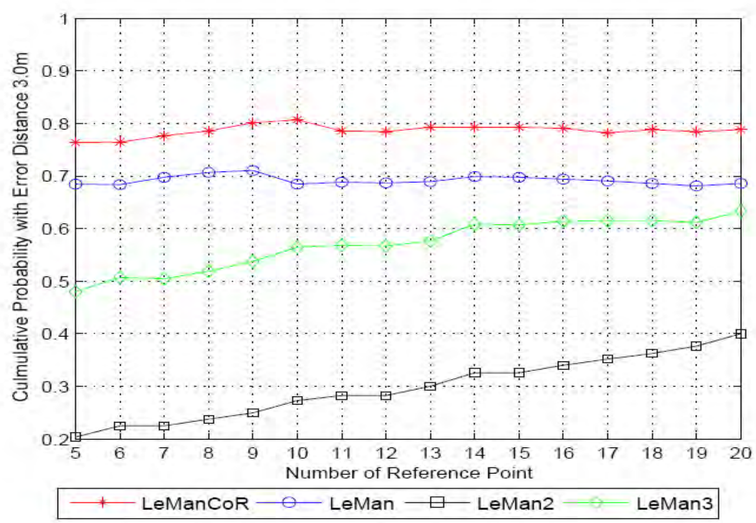




Area: 30 X 40 (81 grids)

Six time periods:

12:30am--01:30am
08:30am--09:30am
12:30pm--01:30pm
04:30pm--05:30pm
08:30pm--09:30pm
10:30pm--11:30pm



LeMan:

Static mapping function learnt from offline data;

LeMan2:

Relearn the mapping function from a few online data

LeMan3:

Combine offline and online data as a whole training data to learn the mapping function.

Domain Transfer Learning: related Works

- **[Huang 06]** reweighting training instances so the training and test means are close in the kernel space.
- **[Raina 06]** learning covariances between features in the source domain to construct an informative prior for the target domain
- **[Lee 07]** sharing a common prior on meta-features between different domains
- **[Smith 07]** proposed a generative classifier based on shifted mixture model to overcome arbitrary sample selection bias

Related Work: Self-taught Learning

[Raina et al. ICML 07] [无师自通]

- Even unlabeled data in target domain can be difficult to obtain.
  - For example, when images of goats and horses are difficult to obtain
- Q: Can we learn with
  - unlabeled images from *other* domains that are easily available +
  - a few labelled images in the target domain

Transfer Learning

Self-taught Learning Overview
[Raina et al. 07]

Input:
- (few) Labeled training data
- Unlabeled data from any classes

Output:
- Predictions of the test data according to the (few) training data

Two steps:
- Applying sparse coding (Ng 2004) algorithm to learn higher-level representation from the unlabeled training data
  - Different distributions and feature space
- Transforming the labeled training data and test data to new representations, and then applying standard classifiers to them.

Learning Higher Level Representation [Raina et al. 07]

- Using the unlabelled data to learn a set of *basis* $b = \{b_1, b_1, \ldots, b_s\}$ and *activations* $a = \{a^{(1)}, a^{(2)}, \ldots, a^{(k)}\}$

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

$$\text{s.t.} \quad \left\| b_j \right\|_2 \le 1, \quad \forall j \in 1, \ldots, s$$

- Achieve Sparse Coding by making *s* far greater than the input dimension and encourage the activation *a* to have low norm, we may obtain large number of high level features.

Examples of Higher Level Features Learned [Raina et al. 07]

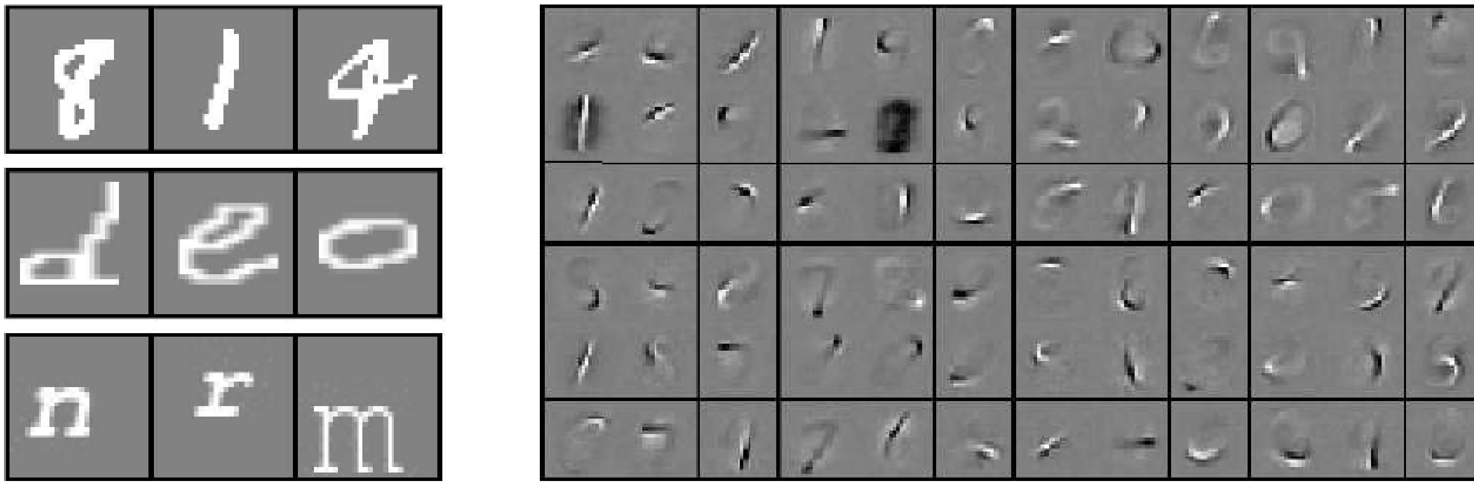


Figure 5. Left: Example images from the handwritten digit dataset (top), the handwritten character dataset (middle) and the font character dataset (bottom). Right: Example sparse coding bases learned on handwritten digits.

Related Work: Multi-task Learning

[Caruana 97]双/多管齐下

- Input: labeled training data for a number of different tasks
- Output: a set of classifiers for all tasks
- Comment:
  - learning the tasks in parallel, using a shared representation
  - multitask learning cares about predicting in many domains.
  - must have same-distribution labeled data from many domains at the same time

Summary of Transfer Learning

- Summarize on two dimensions:
  - Data Requirement:
    - Labeled Source Domain, Unlabeled Target Domain [Dai 07b]
    - Labeled Source Domain, Limited Labeled Target Domain [Dai 07a; Pan 07]
    - Unlabeled Source Domain, Labeled Target Domain [Raina 07]
  - Bridges for Transfer:
    - Feature Based: [Dai 07b; Pan 07; Raina 07]
    - Instance Based: [Dai 07a]
    - Bridge based ...

Conclusions and Future Work: 迁移学习：举一反三

- Transferring the Learned Knowledge
  - Target class can change
  - Training data can change
  - Test data can change
- Future
  - Transfer learning for time sequences
  - Transfer learning for link analysis
  - Transfer learning for clustering

References

- Andrew Arnold (2007) A Comparison of Methods for Transductive Transfer Learning, (unpublished)
- Rich Caruana (1997) Multi-task Learning, in *Machine Learning* (28)
- DARPA Transfer Learning Programme (2005) http://www.darpa.mil/ipto/programs/tl/tl.asp
- W. Dai, Q. Yang, G. Xue and Y. Yu (2007a) Boosting for Transfer Learning. In *Proceedings of ICML 2007*
- W.Dai, G. Xue, Q. Yang and Y. Yu (2007b) Co-clustering based Classification of Out-of-Domain Data. In *Proceedings of SIGKDD 2007*

References

- C. Elkan (2001) The Foundations of Cost-sensitive Learning. In *Proceedings of IJCAI 2001*
- J. Huang, A. Smola, A. Gretton, K. Borgwardt and B. Scholkopf (2006) Correcting Sample Selection Bias by Unlabeled Data. In *Proceedings of NIPS 2006*
- S. Lee, V. Chatalbashev, D. Vickrey and D. Koller (2007) Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks, in *Proceedings of ICML 2007*
- P. Langley (2006) Transfer of Learning in Cognitive System, Invited talk at ICML'06 Workshop on Structural Knowledge Transfer for Machine Learning
- A. Y. Ng (2004) Feature selection, L1 vs. L2 regularization, and rotational invariance, in *Proceedings of ICML 2004*

References

- J. Pan, J. Kwok, Q. Yang and J. Pan (2007) Adaptive Localization in a Dynamic Wifi Environment through Multi-view Learning. In *Proceedings of AAAI 2007*
- R. Raina, A. Ng and D. Koller (2006) Constructing Informative Priors using Transfer Learning. In *Proceedings of ICML 2006*
- R. Raina, A. Battle, H. Lee, B. Packer and A.Y. Ng (2007) Self-taught Learning: Transfer Learning from Unlabelled Data. In *Proceedings of ICML 2007*
- A. Smith and C. Elkan (2007) Making Generative Classifiers Robust to Selection Bias, In *Proceedings of SIGKDD 2007*

References

- G. Widmer and M. Kubat (1996) Learning in the Presence of Concept Drift and Hidden Contexts. In *Machine Learning*
- P. Wu and T. Dietterich (2004) Improving SVM Accuracy by Training on Auxiliary Data Sources. In *Proceedings of ICML 2004*
- B. Zadrozy (2004) Learning and Evaluating Classifiers under Sample Selection Bias. In *Proceedings of ICML 2004*