転移学習のサーベイ

神嶌 敏弘 (産業技術総合研究所)
http://www.kamishima.net/

2009.3.3 AI学会SIG-DMSM研究会

転移学習 (transfer learning)

統一された形式的定義なく，おおまかに次のようなもの

元ドメイン source domain		目標ドメイン target domain
知識データ		
解きたい問題と似てはいるが，**解きたい問題ではない**	⇔	これから**解きたい問題**
解くためのデータや知識は**十分**	⇔	解くためのデータや知識は**不十分**

関連したドメインの知識やデータを転移して
目標ドメインの問題をより高精度で解く

今，転移学習が必要なわけ

例：統計的機械翻訳で，英語を日本語に翻訳する

入力英語文との同時確率を最大にする日本語文を出力

言語モデル

日本語文の事前確率

Pr[日本語]

日本語コーパスから生成
情報化やWebの普及により
膨大な日本語コーパス

急速に改良

×

翻訳モデル

日本語文が与えられたときの
英文の条件付き確率

Pr[英語 | 日本語]

日英の対訳コーパスから生成
増えてはいるが
量は限定されている

改良は限定的

人手による教示データの不足がボトルネック

教示データ不足への対策

能動学習

- より効率的なデータに対して教示情報を選んで与えてもらう
  例：より翻訳に役立ちそうな文を選んで翻訳してもらう
- 任意のデータに対して教示情報がえられる環境が必要

半教師あり学習

- 少量の教師ありデータと大量の教師なしデータ
  例：日本語コーパスと日英の翻訳コーパス
- データの分布に仮定が必要・両方とも目標ドメインのデータ

転移学習

- 目標ドメインとは異なる元ドメインのデータ
  例：日英の翻訳コーパスと日仏や日独の翻訳コーパス

この発表の概要

- **転移学習の現状**：多数の手法が提案され活発な研究

- **形式的定義がなく，体系付けも不十分**

  名称さえ混乱している！：帰納転移 (inductive transfer)，ドメイン適応 (domain adaptation)，マルチタスク学習 (multi-task learning)，knowledge transfer, learning to learn，lifetime learning，共変量シフト (covariate shift)，標本選択バイアス (sample selection bias)

- 二つの文献をもとに体系化を試みる
  - **Daumé:** Hal Daumé III のブログ “natural language processing blog”
  - **Pan&Yang:** S.J.Pan and Q.Yang “A Survey on Transfer Learning”

予稿提出後の考察により，一部予稿と異なる部分がある

転移学習の設定

元と目標ドメインのデータに教示情報(ラベル)あるかどうかで分類

		目標ドメインラベル (Target)	
		あり	**なし**
元ドメインラベル (Source)	**あり**	**(1) S+T+** **帰納転移学習** Inductive Transfer Learning	**(2) S+T-** **トランスダクテイブ転移学習** Transductive Transfer Learning
	なし	**(3) S-T+** **自己教示学習** Self-Taught Learning	**(4) S-T-** **教師なし転移学習** Unsupervised Transfer Learning

※ Pan&Yangでは (1) と(3) を併せて帰納転移学習としている

ドメインの違い

分布の違い　ほとんどの場合

特徴ベクトルとラベルの定義域が等しいとき　$\mathcal{X}^{(S)} = \mathcal{X}^{(T)}$　$\mathcal{Y}^{(S)} = \mathcal{Y}^{(T)}$

データの分布が異なる

$$\Pr[X^{(S)}, Y^{(S)}] \neq \Pr[X^{(T)}, Y^{(T)}]$$

定義域の違い　あまりない

特徴ベクトルの定義域が異なる

$$\mathcal{X}^{(S)} \neq \mathcal{X}^{(T)}$$

ラベルの定義域が異なる

$$\mathcal{Y}^{(S)} \neq \mathcal{Y}^{(T)}$$

※ 上付きの(S)は元ドメイン，(T)は目標ドメインの意味

帰納・トランスダクティブ転移学習

帰納転移学習 (S+T+)

- 元と目標の両方のドメインで，ラベルと特徴ベクトルの対が訓練データとして与えられている
- 同時確率 $\Pr[X^{(S)}, Y^{(S)}] \neq \Pr[X^{(T)}, Y^{(T)}]$ は異なるので，これをうまく一致させるようにする

トランスダクティブ転移学習 (S+T-)

元ドメインにはラベルがあるが，目標ドメインにはない

➡ ラベルの分布をドメイン間で一致させる手がかりがない

➡ $\Pr[Y^{(S)}|X^{(S)}] = \Pr[Y^{(T)}|X^{(T)}]$ を暗黙的に仮定

➡ $\Pr[X^{(S)}] \neq \Pr[X^{(T)}]$ を一致させるように転移する

※ Daumé と Pan&Yang の両者も指摘

自己教示学習 (1)

目標ドメインにはラベルがあるが，元ドメインにはない

特徴ベクトルの分布をドメイン間で一致させる手がかりがない

$\Pr[X^{(S)}] = \Pr[X^{(T)}]$ を暗黙的に仮定 (Daumé の指摘)

統計的機械翻訳の例で，言語の分布と，翻訳の条件付き分布を異なるコーパスから求めるのと同じ？

分布の違いだけなら，転移学習ではなく，
特徴構築の既存の問題と大差ない

ラベルや特徴ベクトルの定義域が違えば
転移学習とみなしてよい？

自己教示学習 (2)

問題設定

- S-T+設定 (元ドメインはラベルなし，目標ドメインラベルあり)
- 元ドメインのデータに与えられるべきラベルの定義域は，目標ドメインのラベルの定義域の上位集合，すなわち $\mathcal{Y}^{(S)} \supseteq \mathcal{Y}^{(T)}$

手法

- **高次表現の獲得**：元ドメインデータを表す，疎な低次元の基底
- **教師なし特徴構築**：目標ドメインの特徴ベクトルだけを，上記の基底を使って表現する
- 教師あり学習：目標ドメインのラベルと，変換後の表現から学習

元ドメインで獲得した高次表現が，目標ドメインには適さない場合も

↓

低次元空間では，暗黙的に $\Pr[X^{(S)}] = \Pr[X^{(T)}]$ を仮定？

R.Raina et. al. “Self-taught Learning: Transfer Learning from Unlabeled Data” ICML2007

教師なし転移学習

目標ドメインにも，元ドメインにもラベルはない

➡ 特徴ベクトルの分布をドメイン間で一致させる手がかりがない

➡ $\Pr[X^{(S)}] = \Pr[X^{(T)}]$ を仮定すると，普通の教師なし学習

(Daumeの指摘)

$\mathcal{X}^{(S)} \neq \mathcal{X}^{(T)}$ の場合に，クラスタリングや次元削減を扱う方法は転移学習と見なせる

異なる特徴空間で，対応付けができるのか？

- 既存手法は，共通する部分空間や，部分的なラベル情報を手かりに
- 部分空間での分布の一致や，低密度部分の対応付けなどでできる？

教師なし学習で，知識を転移して精度は向上する?

- いろいろな仮定をおかないと知識の転移はできないが，本当の精度の向上といえるか？

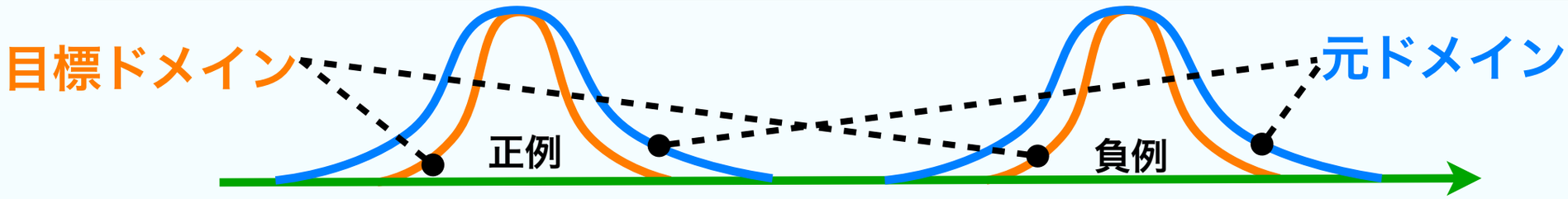
転移仮定と転移モデル

転移仮定　**ドメイン間で何が，どのように似ているのか**

最も一般的な仮定：P[X,Y] が両ドメインで似ている

↓

より詳細な仮定ができれば，より多くの知識を利用可能



この例では，元ドメインの事例はそのまま目標ドメインで利用可能

転移モデル　**転移仮定を，数学的なモデルで表したもの**

同じ転移仮定を異なる転移モデルで表すことも可能

※ Pan&Yang は，転移仮定とモデルに分けず“What to Transfer”によって手法は分類されるとしている．しかし，同じ転移仮定を異なる転移モデルで表すことも可能なので分けるべきと考える．

転移モデルのアプローチ

知識の送信側からのアプローチ

- 入力データを，目標ドメインで使えるように加工する
- 目標ドメインでは，加工されたデータを，通常の手法で学習に利用

事例ベースアプローチ 目標ドメインへの関連性で，事例を重み付け

特徴ベースアプローチ 目標ドメインに合わせて，特徴空間を変換

知識の受信側からのアプローチ

- 元ドメインのデータは，そのまま目標ドメインに送る
- 目標ドメインでは，転移仮定に基づいて利用できる知識を変換しながら利用する方法を採用

モデルベースアプローチ 元ドメインの知識を転移できるモデルの採用

各アプローチの比較

事例・特徴アプローチ

- 転移仮定とは独立に，学習モデルを決定可能
- 転移仮定をモデルには導入できない
- マルチタスク学習には適用できない (?)

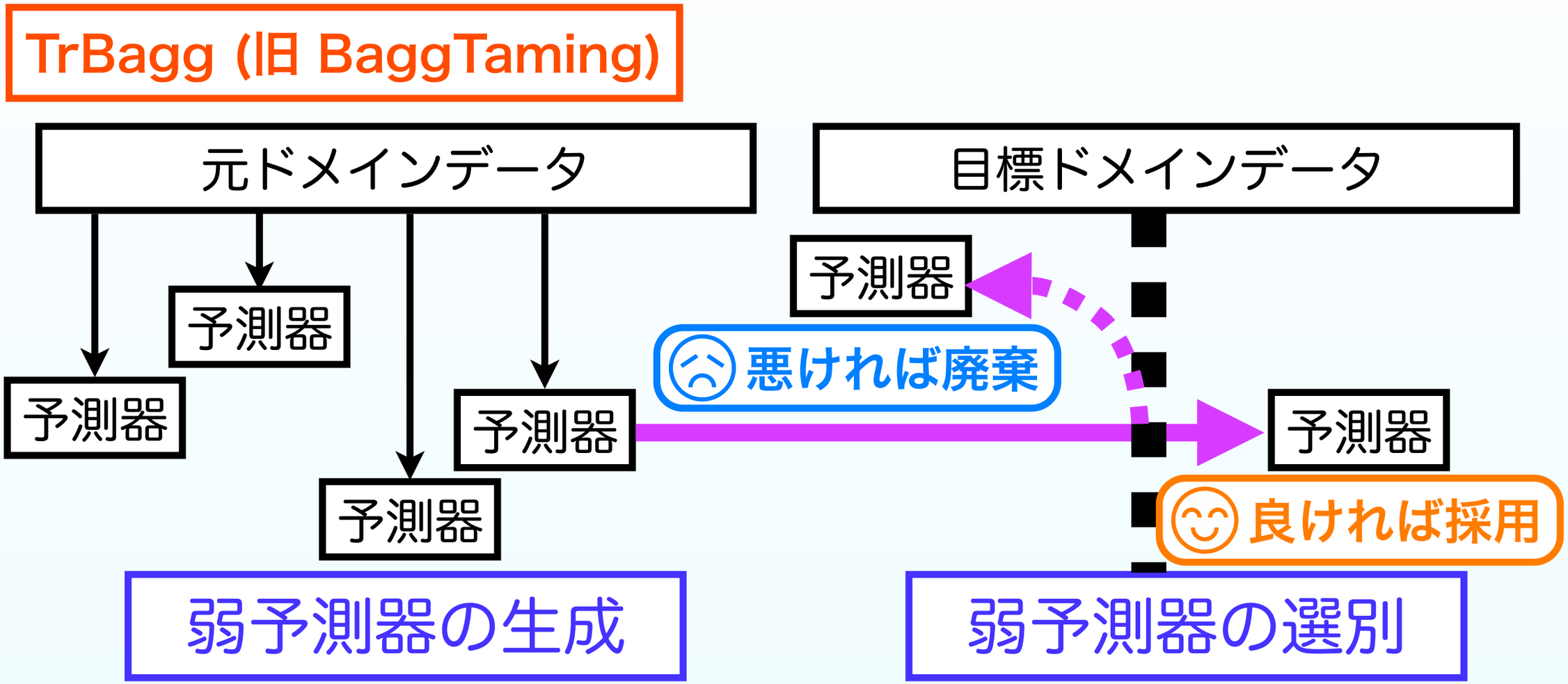
モデルアプローチ

- 学習モデルだけを変えることはできない
- モデルと転移を，より密接に関連付け可能
- マルチタスク学習に適用できる (?)

マルチタスク学習 (Pan & Yang の定義)

- 全ドメインに共通する知識を，全ドメインへ転移，利用することで全てのドメインでの性能向上をめざす
  例：音声認識の話者適応
- 全てのドメインを一度集めないと，それらに共通する知識は分からないので，送信側アプローチは困難 (?)

事例ベースアプローチ



- 元ドメインでブートストラップサンプリングしたデータから弱学習器を学習
- 目標ドメインでの経験誤差を小さくするように，弱予測器を選別
- 最終予測結果は，選別で残った弱予測器の多数決によって決定

神嶌 他 "飼いならし ― 飼育・野生混在データからの学習" AI学会全国大会 (2008)

事例ベースアプローチ

TrAdaBoost

- 転移学習用のAdaBoost
- 目標データは，AdaBoostと同じように，目標データを誤分類したら重みを増やす
- 元データは，誤分類されたら，関係の弱いデータとみなして，重みを減らす
- 全ての分類器ではなく，後半に学習した分類器だけを使う

W.Dai et. al. “Boosting for Transfer Learning” ICML2007

共変量シフト (covariate shift)

- トランスダクティブ転移学習
- $\Pr[Y^{(S)}|X^{(S)}] = \Pr[Y^{(T)}|X^{(T)}]$ を仮定
- $\Pr[X^{(T)}]/\Pr[X^{(S)}]$ で事例を重み付けする

H. Shimodaira “Improving Predictive Inference under Covariate Shift by Weight
ing the Log-Likelihood Function” J. of Statistical Planning and Inference, vol.90 (2000)

特徴ベースアプローチ

類似度学習の利用

- 元ドメインで学習した距離を使い，目標ドメインで再近隣法で分類
- 距離学習は，同じラベルのデータを近づけ，違うラベルのデータを遠ざけるような目的関数をニューラルネットで最小化

S.Thrun “Is Learning The *n*-th Thing Any Easier Than Learning The First?” NIPS1995

いらいらするほど簡単な方法

元ドメイン： $\mathbf{x}^{(S)}$ ➡ $(\mathbf{x}^{(S)}, \mathbf{x}^{(S)}, \mathbf{0})$

目標ドメイン： $\mathbf{x}^{(T)}$ ➡ $(\mathbf{x}^{(T)}, \mathbf{0}, \mathbf{x}^{(T)})$

共通要因　元ドメイン固有　目標ドメイン固有

- 入力データを，長さが3倍の高次元の特徴ベクトルに変換
- 変換後の特徴を用いて，通常の方法で学習

H.Daumeé III “Frustratingly Easy Domain Adaptation” ACL2007

特徴ベースアプローチ

ドメイン間スペクトル分類 (Cross-Domain Spectral Classification)

- 目標ドメインがラベルなしの，トランスダクティブ転移学習
- 次式を最小化することで低次元特徴ベクトル $\mathbf{x}^*$ を得る

データ間の類似度行列

同ラベルのデータを近づけるための罰則項

$$\frac{\mathbf{x}^\top (D - W)\mathbf{x}}{\mathbf{x}^\top D\mathbf{x}} + \beta \|U^\top \mathbf{x}\|^2 + \lambda \frac{\mathbf{x}^\top (D^{(S)} - W^{(S)})\mathbf{x}}{\mathbf{x}^\top D^{(S)}\mathbf{x}}$$

両ドメインに対する項

目標ドメインのみの項

$\mathrm{Diag}(W\mathbf{1})$

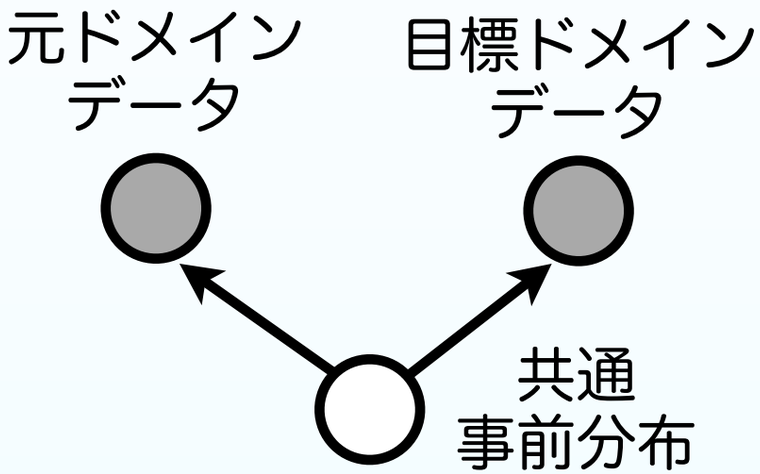
$\mathrm{Diag}(W^{(S)}\mathbf{1})$

※ スペクトラルクラスタリングの Ncut の応用

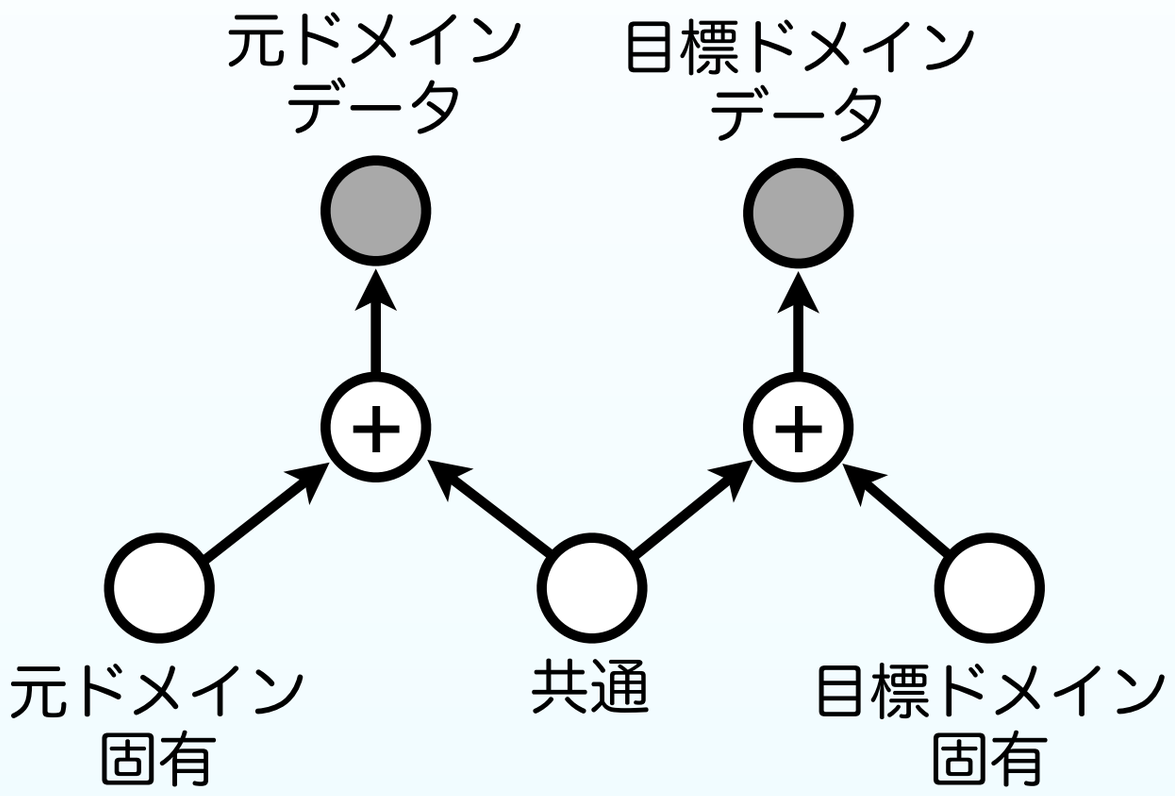
X.Ling et.al. “Spectral Domain-Transfer Learning" KDD2008

モデルベースアプローチ

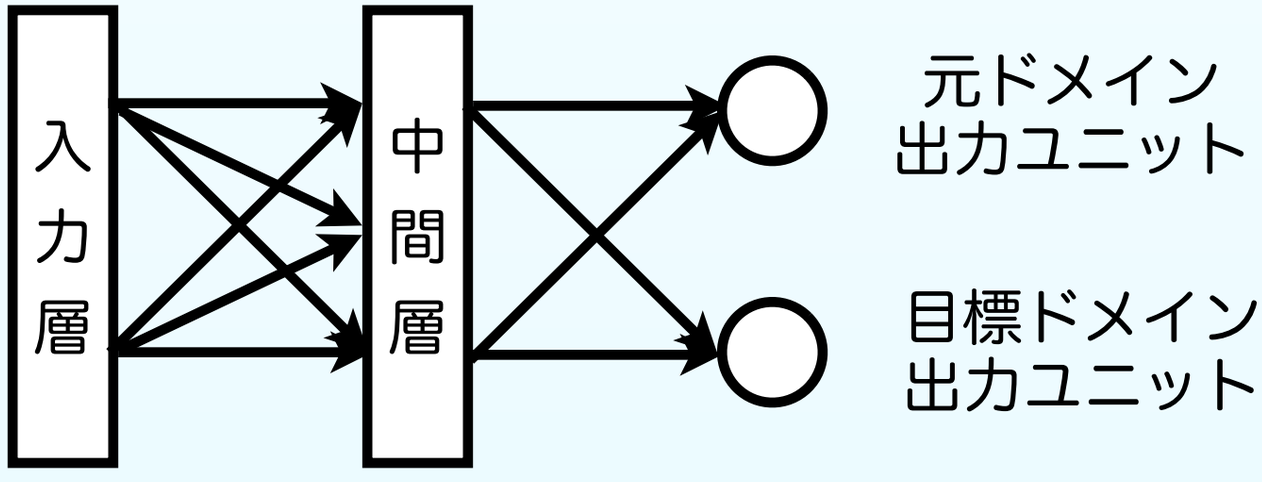
階層ベイズモデル



混合モデル



ニューラルネット



モデルベースアプローチ

Migratory-Logit

- トランスダクティブ転移学習
- 事例の重み付けをモデルベースで行う

事例の有用性を示す重み

目標ドメイン $\Pr[y_i^{(T)}|\mathbf{x}_i^{(T)};\mathbf{w}] = \sigma[y_i^{(T)}\mathbf{w}^\top\mathbf{x}_i^{(T)}]$

元ドメイン $\Pr[y_i^{(S)}|\mathbf{x}_i^{(S)};\mathbf{w},\mu_i] = \sigma[y_i^{(S)}\mathbf{w}^\top\mathbf{x}_i^{(S)} + y_i^{(S)}\mu_i]$

尤度

$$\mathcal{L}[\mathbf{w},\boldsymbol{\mu}|\{(\mathbf{x}_i^{(S)},y_i^{(S)})\},\{\mathbf{x}_i^{(T)}\}] = \sum \ln \Pr[y_i^{(T)}|\mathbf{x}_i^{(T)};\mathbf{w}] + \sum \ln \Pr[y_i^{(S)}|\mathbf{x}_i^{(S)};\mathbf{w},\mu_i]$$

最適化問題 $\max_{\mathbf{w},\boldsymbol{\mu}} \mathcal{L}[\mathbf{w},\boldsymbol{\mu}|\{(\mathbf{x}_i^{(S)},y_i^{(S)})\},\{\mathbf{x}_i^{(T)}\}]$

制約 $\frac{1}{N^{(S)}}\sum y_i^{(S)}\mu_i \leq C, C \geq 0$

$y_i^{(S)}\mu_i \geq 0$

無視される事例の割合

※ 予稿では事例ベースとしていたが，再考察によりモデルベースとする

X.Liao et. al. “Logistic Regression with an Auxiliary Data Streams” ICML2005

まとめ

転移学習：元ドメインの知識を，目標ドメインでの学習に利用する

転移学習の設定：元・目標ドメインのラベルあり・なしで4種類

- 帰納転移学習 (S+T+)，トランスダクテイブ転移学習 (S+T-)
  自己教示学習 (S-T+)，教師なし転移学習 (S-T-)
  - S- の場合については，まだ分からない点も…

転移仮定と転移モデル

- **転移仮定**：両ドメインを結び付けている情報に関する仮定
- **転移モデル**：転移仮定を具体的に数学モデルで表したもの
  ある転移仮定は，送信側と受信側の両方でモデル化できる？

転移学習：負の転移，転移仮定の体系化

おまけ：朱鷺の杜Wiki: http://ibisforest.org/index.php?FrontPage