

## 解説

## 転 移 学 習

## Transfer Learning

神 寫 敏 弘

Toshihiro Kamishima

産業技術総合研究所

National Institute of Advanced Industrial Science and Technology (AIST).

mail@kamishima.net, http://www.kamishima.net/

**Keywords:** transfer learning, inductive transfer, domain adaptation, multitask learning, semi-supervised learning.

## 1. は じ め に

転移学習 (transfer learning) という語は、かなり幅広い機械学習の枠組みに対して使われており、統一された形式的定義を与えることは難しい。だが、形式的ではない定義としては、転移学習のワークショップの論文募集 [TT 05] 中の、次のものが広く受け入れられるだろう。

the problem of retaining and applying the knowledge learned in one or more tasks to efficiently develop an effective hypothesis for a new task

新規タスクの効果的な仮説を効率的に見つけ出すために、一つ以上の別のタスクで学習された知識を得て、それを適用する問題

すなわち、ある問題を効果的かつ効率的に解くために、別の関連した問題のデータや学習結果を再利用するのが転移学習である。

例えば、コンピュータ雑誌の記事中の単語の品詞を特定する問題を考えよう。品詞を分類する規則は、それに含まれる各単語の品詞を示した文書を集めた、コーパスと呼ばれるデータから、機械学習の手法によって獲得する。コンピュータ雑誌記事のコーパスでは、CPU も monitor という単語も名詞であることが多く、これらの単語が現れたら名詞という規則が獲得される。一般に、このコーパス中の文書数が多いと、より正確に品詞を分類できる。ところが、コンピュータ雑誌記事データは全部使ってしまうと、経済関係の新聞記事のコーパスしかないでしょう。それでも、こうした新聞記事も、同じ言語で書かれているので品詞の推定に役立つと予測できるが、単純に二つのコーパスを混ぜたデータから学習すると問題が起きる。例えば、CPU という単語は新聞記事でも名詞として現れるが、monitor という語は『監視する』といった意味で動詞として新聞記事では現れる。こうした、二つのデータの違いをうまく捉え、CPU のような場合だけに新聞記事のデータを利用して、コンピュータ雑誌記事の品詞を特定する規則を獲得したい。このように、関

連しているが異なる部分もあるデータから、目的の問題にも利用できる情報・知識だけを取り込んで、より予測精度の高い規則を得るのが転移学習の目標である。

現在では、Web などから大量のテキストデータはかなり容易に収集できるようになったため、教示情報が不要な教師なし学習で獲得できる言語モデルなどは格段に精度が向上した。一方、音声認識の音響モデルや、文書分類のモデルなどはそれほど向上していない。その理由として、これらのタスクには、教示情報付きの学習データが必要であり、その情報は人手で与えなければならないことがあげられる。しかし、こうした教示情報を大量に与えるのは、費用、人的資源、時間などの制約から、一般には困難だという問題がある。この問題に対処するための機械学習の方針は三つあるだろう。一つ目は、少数のラベルありデータを準備し、多数のラベルなしデータを活用して予測精度を向上させる半教師あり学習 (semi-supervised learning) [Chapelle 06] である。二つ目は、より効率的な分類ができるように選んだ事例に対して、教示情報を与えさせる能動学習 [中村 97, 渡辺 05] である。そして、最後がこの転移学習であり、上記のような需要に応えるため、転移学習に関する研究は活発になっている。

転移学習の考えはかなり以前からあり、1995 年の NIPS のワークショップ [LtL 95] から、機械学習の一分野として認識されるようになった。転移学習の呼び方が定着しつつあるようなので、この呼び方を本稿では採用するが、10 数年にわたっていろいろな呼び方をされてきた。例えば、帰納転移 (inductive transfer)、ドメイン適応 (domain adaptation)、マルチタスク学習 (multitask learning)、knowledge transfer, learning to learn, lifetime learning などである。さらに、共変量シフト (covariate shift) [Shimodaira 00, 杉山 06]、標本選択バイアス [Heckman 79, 星野 09]、音声認識分野の話者適応なども、転移学習のより限定された分野ともみなせる。

Pan & Yang の転移学習のサーベイ [Pan 08b] では、体系的な議論が試みられている。また、自然言語処理で、この転移学習に取り組んでいる Daumé は、ブログ [Daumé III] で、転移学習に対する考えを示している。

本稿では、これらの考えを取り入れつつ、著者自身の考えも含めて、転移学習の分野を2章で俯瞰した後、今までに提案された手法を3章にまとめる。なお、本稿では機械学習に関する用語を用いるが、これらについては、[麻生 03, Bishop 08, 金森 09, 元田 06]などの教科書を参考にされたい。

## 2. 転 移 学 習

転移学習の見方は研究者によりさまざまだが、著者の考えも交えて転移学習を俯瞰する。

### 2.1 表 記

本題に移る前に、本稿での表記をまとめておく。転移する知識の送り手側を元ドメイン (source domain)、受け手側を目標ドメイン (target domain) と呼び、それぞれ、上付きの (S) と (T) で表す。なお、目標ドメインは一つだが、二つ以上の元ドメインから転移する場合もあり得る。いずれかのドメインであることを示すには  $D \in \{S, T\}$  を用いる。各ドメインの入力に該当する確率変数を  $X^{(D)}$  とし、その定義域を  $\mathcal{X}^{(D)}$  とする。この定義域は、 $M$  次元の実数ベクトルである場合が多い。各ドメインに出力が存在するとき、その出力に該当する確率変数を  $Y^{(D)}$  とする。この定義域は  $\mathcal{Y}^{(D)}$  で、二値分類問題なら  $\{0, 1\}$  や  $\{-1, +1\}$  などのラベル集合に、回帰問題なら実数となる。ラベルありの事例は入力と出力の対  $(\mathbf{x}_i^{(D)}, y_i^{(D)})$  で、 $(X^{(D)}, Y^{(D)})$  の実現値である。ラベルなし事例は入力  $\mathbf{x}_i^{(D)}$  のみで、 $X^{(D)}$  の実現値である。各ドメインの訓練事例集合の数は  $N^{(D)}$  で、両ドメインでの総事例数では  $N$  である。各ドメインのデータやラベルを、元データや目標ラベルなどと略す。

### 2.2 転移学習の体系

まず、マルチタスク学習 (multitask learning) について論じる。Daumé は、ドメイン適応・転移学習と、マルチタスク学習とを、それぞれ異なるドメインからの転移と、異なるタスクからの転移と区別している。自然言語処理が専門である Daumé の観点からは、文書分類、固有表現抽出、形態素解析などを異なるタスクとして想定し、訓練に使うコーパスの違いを異なるドメインと想定しているであろう。しかし、ドメインの違いとタスクの違いというのは自明ではないと、著者は考えている。さらに、この分野の研究者の間にも明確なコンセンサスは見当たらないので、本稿では、ドメインとタスクを区別せずに、ドメインという用語を用いる。Pan & Yang は、マルチタスク学習を、両ドメインで同時に学習する転移学習の特殊なものとしている。すなわち、元ドメインと目標ドメインの役割が対称であり、互いに相手のドメインから知識を転移して、両方のドメインで予測精度を向上させることが目的である。本稿では、このマルチタ

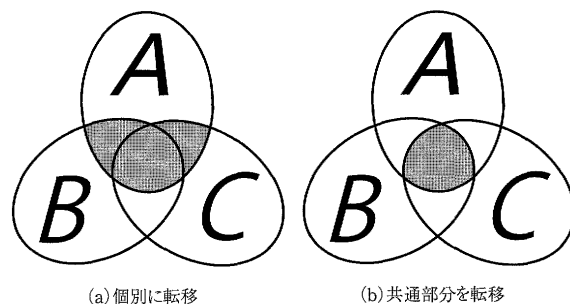


図1 マルチタスク学習で転移される知識

ク学習の定義を採用する。

マルチタスク学習を実現する方法は2種類考えられる。

一つは、各ドメインを目標ドメインとして個別に転移することを、全ドメインについて行う方法である。A, B, Cの三つのドメインがあるとき、最初はBとCを元ドメインに、Aを目標ドメインとして転移学習を適用する。同様に、AとCからBへ、AとBからCへ転移する。もう一つは、全ドメインに共通する知識を求め、それらを全ドメインに転移する。すなわち、A, B, Cに共通する知識を求め、それをを用いてA, B, Cの各ドメインでの学習結果を改善する。これら個別に転移する場合と、共通部分を転移する場合を比較する。転移される知識については、BとCからAに個別に転移する場合、 $A \cap (B \cup C)$  の図1(a)の灰色の部分の知識が転移されるだろう。一方、共通部分を転移する場合、Aには  $A \cap B \cap C$  の図1(b)の部分の知識が転移されるだろう。そのため、個別に転移するほうが多くの知識を転移できることになる。次に、転移にかかるコストを考えると、個別に転移する場合にはすべてのドメインについて転移学習を実行する必要がある、その実行コストは大きい。一方で、共通部分を転移する場合には、共通する知識を求めてしまえば、その知識はどのドメインでも利用でき、そのコストは小さい。さらに、ドメインを追加する場合を考えると、共通部分を転移する場合には、共通する知識はその追加したドメインでも役立つことが多いだろう。音声認識の話者適応では、多くの話者について共通のモデルを事前に作成し、それを個人の少数のデータを用いて、その個人用にモデルを調整する。これは、共通部分を転移するマルチタスク学習とみなせ、100人の話者に共通する知識は101人目にもほぼ適用できるといった利点がある。

次に、メタ学習 (meta learning) [Vilalta 02] との関連について論じる。いわゆる No Free Lunch 定理 [Wolpert 97] により、あらゆる学習手法には何らかのモデルや帰納バイアスが存在する。よって、いかなる学習手法も、あらゆる学習問題において、ほかのすべての学習手法より有効であることはあり得ない。そこで、データの規模・分布や仮説空間などの問題の特性に応じて、適切な学習手法を選択するためのメタ知識を獲得するのがメタ学習である。元ドメインで得られたメタ知識を、それと類似した目標ドメインで利用すると考えれば、転

移学習はメタ学習の部分問題ともみなせる。

### 2.3 転移学習の設定

転移学習は、元・目標の各ドメインの訓練事例に教示情報（出力情報）があるかどうかによって、4 種類の設定が考えられる。Daumé の表記に従ってまとめたのが表 1 である。例えば、(1) の  $S+T+$  の場合は、両ドメインの訓練集合は  $\{(\mathbf{x}_i^{(S)}, y_i^{(S)})\}$  と  $\{(\mathbf{x}_i^{(T)}, y_i^{(T)})\}$  のように、どちらも入力と出力の実現値の対の集合である。Pan & Yang と Daumé の議論を合わせて、本稿では (1) 帰納転移学習 (inductive transfer learning), (2) トランスダクティブ転移学習 (transductive transfer learning), (3) 自己教示学習 (self-taught learning), そして (4) 教師なし転移学習 (unsupervised transfer learning) と呼ぶことにする。以下、各学習問題を順に論じていこう。

表 1 教示情報の有無による転移学習の分類

		目標ドメインラベル	
		あり	なし
元ドメイン ラベル	あり	(1) $S+T+$	(2) $S+T-$
	なし	(3) $S-T+$	(4) $S-T-$

(1)  $S+T+$  : 帰納転移学習の研究は最も多く、標準的な転移学習の標準設定といえる。ドメインの違いを、分布の違い  $\Pr^{(S)}[X, Y] \neq \Pr^{(T)}[X, Y]$  と考えるものが多い。だが、一部には、定義域の違い  $\mathcal{X}^{(S)} \neq \mathcal{X}^{(T)}$  や  $\mathcal{Y}^{(S)} \neq \mathcal{Y}^{(T)}$  を考えるものもある。例えば、自然言語処理において、両ドメインの語彙集合の一部に違いがある場合などである。

(2)  $S+T-$  : トランスダクティブ転移学習は帰納転移学習の次に研究が多い。ラベルのない目標ドメインのデータの適切なラベルを予測することが目的である。Daumé や Pan & Yang もともに指摘することだが、目標ドメインのラベルがなくラベルの分布を一致させる手掛りがないので、 $\Pr^{(S)}[Y|X] = \Pr^{(T)}[Y|X]$  を暗黙的、もしくは明示的に仮定している。そして、 $\Pr^{(S)}[X]$  と  $\Pr^{(T)}[X]$  をうまく一致させることで知識の転移を目指すことになる。なお、半教師あり学習との違いは、ラベルありとなしのデータで入力  $X$  の分布が異なることである。実際に、半教師あり学習手法も、転移学習アルゴリズムの論文では比較実験されているが、転移学習手法のほうがほとんどの場合、優れている。

(3)  $S-T+$  : この自己教示学習については議論すべき点があると著者は考える。Daumé はこの種の転移学習については、 $\Pr^{(S)}[X] = \Pr^{(T)}[X]$  の仮定が必要だと述べている。この仮定があれば、バイズ型の学習で従来からある、事前分布と尤度関数を異なる訓練集合から学習する場合に該当する。この種の問題を最初に提起した Raina らの自己教示学習 [Raina 07] では、これとは異なる主張をしている。この方法では、元ドメインデータを使って特徴の疎な部分空間を求め、その部分空間を使っ

て目標ドメインのラベルありデータから学習する。このとき、 $\Pr^{(S)}[X] = \Pr^{(T)}[X]$  を仮定しないこと、および目標ドメインのラベルの定義域には含まれないラベルが付けられるべきデータが元ドメインにはある、すなわち、 $\mathcal{Y}^{(S)} \supset \mathcal{Y}^{(T)}$  であることを仮定する。Raina の提案する枠組みが、転移学習であるかについて論じたい。 $\Pr^{(S)}[X] \neq \Pr^{(T)}[X]$  であるかを実データで示すのは困難なため、Raina らに明確な検証を要求するのはやや過剰ではある。しかし、提案手法の元ドメインでの次元削減は、一般の教師なし次元削減であり、特に目標ドメインへの適応を考慮していないように思える。また、後者のラベル集合についても、やはり元ドメインにラベル情報がないため特別な操作は行われていない。よって、提案手法が予測精度の向上に効果的であることは実験により明確に示されているが、 $\Pr^{(S)}[X] \neq \Pr^{(T)}[X]$  という仮定が実験の状況で成立していることの検証は不十分だと著者には思える。

(4)  $S-T-$  : この場合でも  $\Pr^{(S)}[X] = \Pr^{(T)}[X]$  の仮定が必要で、通常の教師なし学習と同じになると、Daumé と同様に著者も考えている。ただし、定義域が異なる場合、すなわち、 $\mathcal{X}^{(S)} \neq \mathcal{X}^{(T)}$  であれば、新たな枠組みとみなせるかもしれない。[Wang 08] など、いくつかの研究が提案され始めた時期なので、その位置づけは今後とも考察を続けたい。

### 2.4 転移仮定と転移モデル

前節では、データの分布や定義域が両ドメインで、単に異なっているかどうかについて論じたが、ここでは、どの程度や、どの点が異なっているかについて論じる。元ドメインでの分布や定義域が全く異なれば、目標ドメインでは全く役立たないであろうし、逆にわずかな差しかなければ特別な転移をする必要はないだろう。このように転移学習は、適度にドメインや分布が『似ている』データが元データであることを想定しているが、この『似ている』ことを形式的に定義することは難しい。これは、転移学習研究の黎明期から最も重要な課題として認識されているが [Caruana 97], いまだに明確な解は示されていない。そこで、普遍的な定義を与えるよりも、どんな観点で類似していて、どのように知識を転移できるかを仮定し、その仮定をどのように数学的モデルで表したかを、それぞれの学習手法で明確にすることが必要だと著者は考えている。これらを、転移仮定と転移モデルとそれぞれここでは呼ぼう。適用した転移学習手法の、転移仮定と転移モデルが、二つのドメイン間で適切ならば知識の転移は成功し、不適切ならかえって悪くなる『負の転移 (negative transfer) [Rosenstein 05]』が生じるであろう。

まず、転移仮定について考察する。理論系の研究で広く導入されている転移仮定に、二つのドメインの分布間のダイバージェンスで測り、それが十分に小さいとする

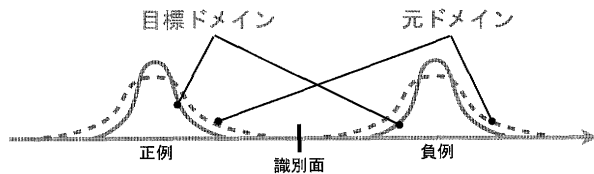
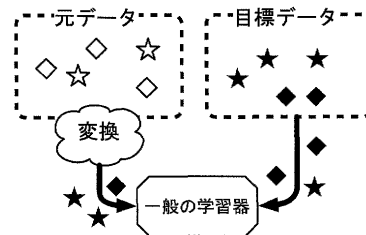


図2 簡単に転移ができる分布の例

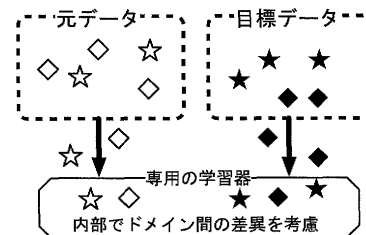
ものがある。ここで、図2のような例を考えてみよう。正と負クラスの目標データが、それぞれ、別のガウス分布から生成されているとする。元データもやはりガウス分布に従うのだが、その分布の中心は目標データのそれと等しいが、分散はずっと大きいとしよう。この場合、決定平面は不変なので、そのまま元ドメインのデータや知識はすべて転移できるだろう。しかし、これら両ドメインの分布は一致していないので、ダイバージェンスに基づく転移仮定では、元ドメイン知識の一部は活用されないだろう。この例のように、ドメインがどの点で類似していて、どの点で相違しているかについてより多くの情報があれば、それを活用しドメイン間に類似性を考えることができる。広く適用できるダイバージェンスに基づくような転移仮定はもちろん重要である。しかし、問題から転移できる知識を明確にし、それに合わせたいろいろな転移仮定に基づく手法を選択できるように、研究が進められるべきだろう。

次に、転移仮定を数学的にモデル化した転移モデルについて論じる。Pan & Yang, Daumé, および著者の考えを合わせて、このモデル化のアプローチを、データのどの部分を転移時に変換するかによる分類と、転移のどの段階でデータを変換するかによる分類の二つの軸で分類する。一つ目の、変換する部分による分類には特徴ベース (feature-based) と事例ベース (instance-based) がある。特徴ベースアプローチでは、元ドメインの特徴から目標ドメインでも有用なものを選択したり、元ドメインの特徴空間の中で目標ドメインでも有用な部分空間を求めたりする。事例ベースアプローチでは、元ドメインの事例を、目標ドメインでの有用性に基づいて重み付けしたり、選択したりする。

データを変換する段階の違いによる分類では、分離アプローチ (separated) と統合アプローチ (integrated) とがある。図3 (a) の分離アプローチでは、元ドメイン側でデータや知識を、目標ドメインの問題に適合するように変換する。目標ドメインでは、変換されたデータを目標ドメインのデータとして扱い、一般の学習手法で解く。もう一方の統合アプローチ (図3 (b)) では、元ドメインのデータは変換されずに目標ドメインに渡される。そして、元ドメインと目標ドメインの差異を配慮したモデルや手続きによって、目標問題に適した学習結果を得る。分離アプローチでは、学習手法と転移仮定は独立に選ぶことができる利点がある。統合アプローチでは、転移仮定に基づいた変換がモデルに組み込まれているた



(a) 分離アプローチ



(b) 統合アプローチ

図3 データを変換する段階の相違に基づく転移モデルの分類

め学習手法と切り離すのは困難となる。その一方で、モデルと転移仮定をより密接に結びつけることが可能になることが利点である。

### 3. 転移学習手法

本章では、転移学習の手法をいくつか紹介する。

#### 3.1 全 般

これまでに述べたように、転移学習全般についての議論としては [麻生 09, Daumé III, Pan 08b] がある。専門のワークショップとしては [TM 09, TSL 09, TT 05] などがある。転移学習に関連したアイディアは 90 年代中頃から提案されてきてはいた [Caruana 96, Munro 97]。だが、この種の問題が機械学習の研究分野の一つとして意識されたのは次の二つの論文の影響が大きいだろう。

一つ目は、Thrun の [Thrun 96] である。ある画像の集合中で、ある特定の画像であるかどうかを識別する問題において、ほかの画像集合を活用する帰納転移学習問題を論じた。特徴ベース・分離型の方法であり、第1段階として元ドメインの画像集合で同じカテゴリーの画像は近く、違うものは遠くなるような目的関数を定義し、ニューラルネットでも最適化することで特徴空間の変換を求める。その後、距離減衰付きの  $k$  近隣法と、目標データのラベルに加えて、変換した距離空間から推定した勾配も同時に学習させる explanation-based ニューラルネットの二つの方法で識別する方法を提案している。

Caruana の [Caruana 97] は、複数のタスクを共通の中間層で学習するニューラルネット (特徴ベース・統合型) を、いくつかの帰納転移学習問題に適用し、転移学習について論考した。この時点以前の研究を俯瞰した論文であり、一読されることを薦める。転移学習の有効性

を三つのデータで実験的に示したのち、転移が成功する理由、転移が可能である状況、ニューラルネット以外の手法の適用可能性について考察している。ここでは、転移が成功する四つの理由を紹介しよう。(1) データ拡張 (data amplification): 共通する因子がある複数タスクがあれば、共通因子のノイズがタスク間でキャンセルされて緩和されることで学習が容易になる。(2) 属性選択 (attribute selection): 複数のタスクを同時に学習させることで、それらのタスクで共通に有用な特徴をより容易に発見できるようになる。(3) 立ち聞き (eavesdropping): タスク間で学習の難易度に差があれば、困難なタスクは容易なタスクから情報を得てより簡単に学習が可能になる。(4) 表現バイアス (representation bias): 各タスクに複数の局所解があるとき、複数のタスクに共通な局所解が選ばれやすくなる。逆に、ほかのタスクで使われない局所解は使われにくくなる。

以降、特徴・事例ベースと、分離・統合型の 4 種類の組合せについて、いくつかの方法を紹介する。

### 3.2 特徴ベース・分離型

[Daumé 07] は非常に簡便な帰納転移学習用の方法を提案している。入力特徴ベクトルや 0 ベクトルを連結して、共通、目標ドメイン、元ドメインに対応した三つの部分で構成される、元の 3 倍の長さのベクトルを生成する。具体的に、目標ドメインの事例 ( $\mathbf{x}^{(T)}, \mathbf{y}^{(T)}$ ) は、( $\langle \mathbf{x}^{(T)}, \mathbf{0}, \mathbf{x}^{(T)} \rangle, \mathbf{y}^{(T)}$ ) に、元ドメインの事例 ( $\mathbf{x}^{(S)}, \mathbf{y}^{(S)}$ ) は、( $\langle \mathbf{x}^{(S)}, \mathbf{x}^{(S)}, \mathbf{0} \rangle, \mathbf{y}^{(S)}$ ) に変換したのち、通常のカテゴリ問題として解く。すると、両ドメインに共通して利用される特徴は、三つのうちで、両ドメインで特徴が共通している最初の部分が分類規則で利用される。一方、両ドメインで働きが異なる特徴は、一方のドメインの値が 0 になっていて両ドメインで異なっている残りの部分が自然に利用されるようになる。

### 3.3 特徴ベース・統合型

最初は、転移学習の枠組みを組み込んだ生成モデルを紹介する。3.1 節で紹介した [Caruana 97] のニューラルネットなども、この種の手法とみなせるだろう。次の基本的な階層ベイズモデルはマルチタスク学習を目的としているといえる。

$$\prod_i^K \Pr[\mathbf{x}_i, y_i | \theta_i] \Pr[\theta_i | \Psi]$$

例えば、病院  $i$  でのデータは、各病院の固有の要因を反映したパラメータ  $\theta_i$  に基づくモデルで記述される。そして、各病院に固有の要因は、全病院に共通する要因  $\Psi$  に基づくモデルで記述されている。このようなモデルは明らかにマルチタスク学習を念頭に置いており、こうした手法として [Raina 06] などがあげられる。

階層的なモデルではなく、混合分布を使ったモデルとして [Daumé 06] がある。これは、ラベルの分布が、特

徴関数のベクトル  $\mathbf{f}$  のギブズ分布  $\exp(\lambda^T \mathbf{f})$  に比例する最大エントロピーモデルを拡張したもので、帰納転移学習を目的としている。元データの生成過程もほぼ同様なので、目標データのみについて述べる。目標ドメインと共通ドメインのいずれかのモデルを選択する潜在変数を導入する。この潜在変数が選んだ目標・共通いずれかのモデルにより、特徴ベクトルの各要素の値が決まる。同様にパラメータ  $\lambda$  もこの潜在変数によって選ばれる。こうして選ばれた特徴ベクトルとパラメータで決まるギブズ分布に従い、目標データは生成される。

[Xue 08] のモデルは、pLSA [Hofmann 99] に基づく方法で、両ドメインの文書は共通する語彙集合に依存する点では階層的だが、両ドメインの影響を調整する重みがある点は混合型ともみなせ、階層型と混合型の中間的な方法とみなせる。さらに、元ドメインのラベルを目標ドメインに反映させる must/cannot リンク [Wagstaff 01] を導入してトランスダクティブ転移学習を実現している。[Ando 08] も同様に両ドメインに共通の因子をもたせるタイプのもので、情報ボトルネック [Tishby 99] の枠組みを転移学習に拡張したものである。

[Dai 07a] は、二つの集合を共通する要因に基づいて分割する共クラスタリング (co-clustering) を利用する。トランスダクティブ転移学習での文書分類で、元ドメインの文書集合のカテゴリと、それを記述する語彙集合の分割を関連付けると同時に、その語彙集合分割と関連付けた目標文書集合の分割を獲得することで、目標ドメインでの分類を行う。

元ドメインのデータのうち、目標ドメインでも有用な、特徴の部分空間を見つけ出し、その部分空間内に射影した元データを目標ドメインで利用する方法を紹介する。

[Argyriou 07] は、複数のタスクを同時に解くマルチタスク学習である。 $K$  種類のタスクそれぞれについて、 $k$  番目の訓練集合から予測関数  $f_k(\mathbf{x}) = \sum_m^M a_{mk} (\mathbf{u}_m^T \mathbf{x})$  を学習する。ただし、 $M$  は特徴数、 $\mathbf{u}_i$  は直交な射影。ほとんどのタスクにおいて係数  $a_{mk}$  は 0 であるようにしたい。そこで、 $u_i$  と  $a_{mk}$  をまとめた行列  $\mathbf{U}$  と  $\mathbf{A}$  を導入し、次の目的関数の最小化を考える：

$$\sum_{k=1}^K \sum_{i=1}^N \mathcal{L}(y_{ki}, \mathbf{a}_k^T (\mathbf{U}^T \mathbf{x}_{ki})) + \gamma \|\mathbf{A}\|_2^2$$

記号の詳細は省くが、第 1 項は全タスク、全データにわたる損失関数の総和であり、データによく当てはまるようにする。第 2 項は正則化項で、全タスクについて  $L_2$  ノルムを各特徴ごとに求め、それらの  $L_1$  ノルムを求める。この正則化項により、全タスクにわたって不要な特徴の重みは 0 となり、全タスクに共通に有用な特徴だけが選択されることになる。

[Ling 08] はスペクトルクラスタリング基準の一つである normalized cut [Shi 00] を用いたトランスダクティブ転移学習である。 $\mathbf{W}$  と  $\mathbf{W}^{(S)}$  をそれぞれ全データと元データ間の類似度行列、また  $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{1})$ 、 $\mathbf{D}^{(S)} = \text{diag}(\mathbf{W}^{(S)}\mathbf{1})$ 、

として、次式を最小化する  $\mathbf{x}$  を見つける。

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

第1項はスペクトラルクラスタリングのレイリー商である。この式の最小化により、類似度行列  $\mathbf{W}$  に基づき二つのクラスタに分けると、各データが0と1のどちらのクラスタに近いかがベクトル  $\mathbf{x}$  で示される。第2項は、詳細は省くが、元データで同じラベルのデータが同じクラスタになりやすいようにする制約項である。第3項は、第1項と似ているが、元データのみを分割している。これら三つの項の影響をパラメータ  $\beta$  と  $\lambda$  で調節することで、元ドメインのラベルと、データ間の類似度を反映した、目標ドメインの分割が得られる。

2・3節でも述べた [Raina 07] の自己教示学習も特徴の部分空間を利用する。元ドメインのラベルなしデータから、 $m$  個の  $M$  次元基底ベクトル  $\mathbf{b}_1, \dots, \mathbf{b}_m$  を次式の最小化で求める。

$$\min_{\mathbf{a}, \mathbf{b}} \sum_i \|\mathbf{x}_i^{(s)} - \sum_j \mathbf{a}_{ij} \mathbf{b}_j\|_2^2 + \beta \|\mathbf{a}_i\|_1, \text{ s.t. } \|\mathbf{b}_j\|_2 \leq 1$$

ここで、 $\mathbf{a}_{ij}$  は  $i$  番目の事例についての、 $m$  次元部分空間中の  $j$  番目の特徴に対する係数である。こうして求めた基底  $\mathbf{b}$  を用いて、次式で目標ドメインの特徴データを射影する。

$$\arg \min_{\mathbf{c}_i} \|\mathbf{x}_i^{(t)} - \sum_j \mathbf{c}_{ij} \mathbf{b}_j\|_2^2 + \gamma \|\mathbf{c}_i\|_1$$

こうして求めた  $m$  次元に射影した特徴ベクトル  $\mathbf{c}_i$  と、元のラベル  $y_i^{(t)}$  を用いて学習をする。

次元削減を使うほかの方法をいくつかあげておく。[Ando 05] は、特定の特徴の状態をほかの特徴から予測するといった目標問題に関連した補助問題を利用して、目標問題を解くのに適切な部分空間を獲得する。[Argyriou 08] は、類似したタスクのクラスタを求め、各クラスタごとに部分空間を求める。[Satpal 07] は、特徴の各ドメインでの期待値を求め、それらの値のドメイン間距離で特徴を重み付けするトランスダクティブ帰納転移である。[Wang 08] は次元削減そのものを目標にしておき、目標ドメインのデータをクラスタリングでラベル付けするステップと、両ドメインのラベルを用いて Fisher 判別分析による次元削減を繰り返すトランスダクティブ転移学習である。ほかに [Do 06, Pan 08a] などとも次元削減を用いる方法とみなせるだろう。

カーネルを使った特徴ベースの手法である [Rückert 08] を紹介する。複数の元ドメインがある状況を想定し、基礎カーネルの線形結合で決まるカーネルの重みを、各元ドメインの問題に最適化する。そして、各元データとそれに最適なカーネルの対を訓練事例として、任意のデータに対して適切なカーネルを出力する関数をメタ学習する。このカーネルは、各元データと目標データとの間に定義されたヒストグラムカーネルと、各元ドメインの最適カーネルとの線形結合としてモデル化されている。メタ学習により学習

した関数に、目標データを代入すれば、目標ドメインの問題を解くのに適切なカーネルが得られる。

### 3・4 事例ベース・分離型

単独では予測精度の低い弱分類器を組み合わせて予測性能を向上させるアンサンブル学習に基づく方法をいくつか示す。これらは、弱学習器として通常の学習器を用いるため、分離型である。

代表的なブースティングである AdaBoost [Freund 96, フロインド 99] を、帰納転移学習問題に拡張したのが TrAdaBoost [Dai 07b] である。通常のブースティングでは訓練事例を重み付けし、弱学習器はこの重みを反映した誤差を最小化するように学習する。その弱分類器に事例を分類させて、分類を誤った、すなわち苦手な事例を重視するように、事例の重みを更新する。このように学習した弱分類器の予測を、分類誤差に応じて凝集することで、最終的な予測結果を得る。TrAdaBoost では、苦手な目標ドメインのデータの重みを増やす点は同じだが、誤って分類した元ドメインデータは目標問題への関連が弱いとみなして、その重みを下げる。さらに、すべての弱分類器ではなく、一連の弱分類器のうち後半で獲得されたもののみを採用する点も異なる。もう少し形式的に述べると、弱分類器を学習する総回数を  $T$ 、 $t$  回目の学習器のラベル出力を  $h_t(\mathbf{x}_i) \in \{0, 1\}$ 、元と目標の両訓練データに対する  $h_t(\cdot)$  の重み付き誤差を  $\varepsilon_t < 1/2$  とする。 $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$  として、誤分類された目標データのみ、その重みを  $1/\beta_t$  倍する。元データも、分類を誤ったときにのみ  $1/(1 + \sqrt{(2 \ln t)/T})$  倍する。最終の分類は次式を満たすときラベル 1 を、それ以外でラベル 0 を出力する。

$$\prod_{t=\lceil T/2 \rceil}^T \beta_t^{-h_t(\mathbf{x})} \geq \prod_{t=\lceil T/2 \rceil}^T \beta_t^{-1/2}$$

順位付け用の RankBoost [Freund 03] にこのアイデアを導入した [数原 09] などもある。

もう一つのアンサンブル学習であるバギング [Breiman 96] を転移学習用に拡張したのが TrBagg [Kamishima 09] である。バギングは、訓練データからブートストラップサンプリング、すなわち重複を許してサンプリングして生成した訓練事例集合をいくつかつくり、それぞれから弱分類器をつくる。最終分類結果はこれらの弱分類器の多数決で決める。TrBagg は目標データからではなく、両データを併せたデータからサンプリングして弱分類器を学習する。すると、これらの弱学習器には目標ドメインに適したものとそうでないものが混在してしまう。そこで、目標データを分類し、その経験誤差を悪化させるような弱分類器を廃棄し、残った弱学習器の多数決により最終結果を決める。

マルチタスク学習で、複数の元データから学習された分類器を統合する方法がいくつかある。[Eaton 08] は、

目標データのみで学習した場合と、元データを考慮して学習した場合での予測精度の差で定義される転移可能性の概念を導入している。最終的な分類器は、この転移可能性の大きな分類器が、より重視されるようにつくられる。トランスダクティブ転移学習の [Gao 08] は、各元データから訓練した分類器の重要性を入力データの一致性に基づいて決める。この一致性は、クラス境界が分布の低密度部分にあるという仮定のもと、目標データのクラスタリング結果と、元データのラベル付けの一致性で測る。

共変量シフト (covariate shift) [Shimodaira 00] は転移学習とは異なる経緯で提案されたが、トランスダクティブ転移学習と類似している。この問題での仮定は、教師情報ありの元データの入力分布は  $\Pr^{(S)}[X]$  から得られるが、予測の対象となる目標データの入力分布は  $\Pr^{(T)}[X]$  に従う。そして、入力が与えられたときの出力の分布は両ドメインで等しい、すなわち  $\Pr^{(S)}[Y|X] = \Pr^{(T)}[Y|X]$  である。出力が実数である回帰問題において、両ドメインの入力分布は既知としたとき、次式を最小化するようにパラメータ  $\theta$  を定めることで、漸近的に目標ドメインで適切な予測ができる。

$$\sum_i \frac{\Pr^{(T)}[\mathbf{x}_i]}{\Pr^{(S)}[\mathbf{x}_i]} \text{loss}(y_i^{(S)}, \mathbf{x}_i^{(S)}; \theta)$$

共変量シフトについては [杉山 06] の解説が詳しい。その他、[Huang 07, 杉山 07a, Sugiyama 07b] などの研究もある。

調査対象である母集団の特徴が、実際に標本を集めた集団とは異なる標本選択バイアス (sample selection bias) [Heckman 79, 星野 09] も転移学習と関連がある。機械学習分野で標本選択バイアスを扱った [Zadrozny 04] では、入力  $\mathbf{x}$  と出力  $y$  に加えて隠れ変数  $s \in \{0, 1\}$  を導入する。標本  $(\mathbf{x}, y)$  はランダムに母集団からサンプリングされているのだが、隠れ変数が  $s=1$  であれば標本は参照できて分析に利用できるが、 $s=0$  であれば参照できない。ここでは、 $\mathbf{x}$  が与えられたとき  $s$  と  $y$  が条件付き独立である、すなわち  $\Pr[y|s, \mathbf{x}] = \Pr[y|\mathbf{x}]$  である場合について次の結果を示した。訓練事例が増えたときに学習器の出力が  $\Pr[y|\mathbf{x}]$  のみに依存する局所学習器と、 $\Pr[y|\mathbf{x}]$  と  $\Pr[\mathbf{x}]$  の両方に依存する大域的学習器とを考えると、局所的学習器は標本選択バイアスに影響されないが、大域的学習器は影響を受ける。なぜなら、 $\Pr[y|s, \mathbf{x}] = \Pr[y|\mathbf{x}]$  より  $\Pr[y|\mathbf{x}]$  は  $s$  に依存せず、そのままデータから推定できるが、 $\Pr[\mathbf{x}]$  は  $s$  に依存しているからである。独立性仮定のないベイズ分類器、ロジスティック回帰、ハードマージンの SVM などは局所的で、単純ベイズ、決定木、ソフトマージンの SVM は大域的学習器である。

トランスダクティブ転移学習の [Xing 07] では、元データで学習した分類器で目標データを分類する。ここで、元データと目標データには違いがあるため、一度、両者の中間である両方のデータを混ぜたデータへ転移し、そ

の分類結果をさらに目標データへの転移に用いるという bridged refinement を提案している。

### 3.5 事例ベース・統合型

Migratory-Logit [Liao 05] は、帰納転移学習問題を扱う。ロジスティック回帰をベースにした次の目的関数を最適化するように、各元データの重要度  $\mu_i$  と、重み  $\mathbf{w}$  とを求める。

$$\begin{aligned} \max_{\mathbf{w}, \mu} \sum_i \sigma(y_i^{(T)} \mathbf{w}^T \mathbf{x}_i^{(T)}) + \sum_i \ln \sigma(y_i^{(S)} \mathbf{w}^T \mathbf{x}_i^{(S)} + y_i^{(S)} \mu_i) \\ \text{subject to } \frac{1}{N^{(S)}} \sum_i y_i^{(S)} \mu_i \leq C, \quad C \geq 0, y_i^{(S)} \mu_i \geq 0 \end{aligned}$$

ここで、ラベル変数は  $y_i^{(D)} \in \{-1, +1\}$  なので、 $y_i \mathbf{w}^T \mathbf{x}_i$  が正ならば予測は正しく、負なら誤りである。目標データについては通常のロジスティック回帰だが、元データには  $y_i^{(S)} \mu_i^{(S)}$  がある。元データ  $\mathbf{x}_i^{(S)}$  が誤分類されると  $y_i^{(S)} \mathbf{w}^T \mathbf{x}_i^{(S)}$  は負になるが、この正の追加項  $y_i^{(S)} \mu_i^{(S)}$  によって『埋め合わせる』、すなわち一部の元データを無視する。ところが、無制限に元データを無視してしまうと、当然ながら元データは役に立たないので、この追加項の総和をただだか  $N^{(S)}C$  に抑える。よって、このパラメータ  $C$  を大きくすると、より多数の元データが無視されることになる。

より簡潔な定式化として [Wu 04] のような方法もある。これは、元データと目標データそれぞれに対する損失の線形和を全体の損失とし、元データに対する損失が、目標データのそれより相対的に小さく評価されるように係数を掛けている。この方法は、3.3 節で紹介した混合型の生成モデルと類似しているが、元・目標モデルに共通する因子があるため、本稿では特徴ベースに分類した。生成モデルを使う学習手法でも、両ドメインをつなぐ要素がなければ、両ドメインからの訓練事例を重み付けしたものとみなせるため、[Wu 04] の方法は事例ベースのアプローチとして扱った。

### 3.6 その他の観点

最後に、手法そのものの以外で、転移学習に関連する話題を取り上げる。

カリキュラム学習 (curriculum learning) [Bengio 09] は、目標タスクを解くための訓練集合を最初から与えるのではなく、目的タスクとは関連があるが、それよりも相対的に容易なタスクの事例を先に与えるものである。例えば、任意形状の丸・三角・四角の 3 種類の画像を識別する目標タスクに対して、正円・正三角形・正方形の 3 種類だけを含むような訓練事例を初期的に与える。また、自然言語処理での学習で、当初は語彙数を減らして訓練する場合などもある。こうしたカリキュラム学習の枠組みを採用することで、多峰性の目的関数をオンライン学習するとき、良い解に早く収束するだけでなく、汎化性能も良くなる事例を実験的に示している。



[Rosenstein 05] は、転移学習で元データを利用したことで、かえって予測精度が低下する負の転移 (negative transfer) 現象について論じた。元データと目標データが共通の事前分布に依存する階層ベイズモデルにおいて、元と目標データの分布の差異が大きいと負の転移が起きることを実験的に示している。

[Shi 08] では、こうした負の転移の問題に対処するため、元データからの転移に、人手によるラベル付けを加える枠組みを提案している。まず、転移学習器の予測が信頼できればそのラベルを、信頼できなければ専門家にラベル付けさせる。この手続きで、得られたラベル付き目標データから最終の分類器を学習する。転移学習による予測の信頼性は、元データと目標データそれぞれで訓練した分類器の予測の一致、ラベルのある目標データ数、分類器の確信度などに基づいて判定している。

目標ドメインでの誤差限界についての議論もある。元ドメインや目標ドメインで学習した予測器の線形結合で目標問題を解く場合に、汎化誤差を最小化するように重みを定める問題を扱う [Blitzer 08, Crammer 08] や、転移学習で獲得した予測器の汎化誤差について論じた [Ben-David 07, Dai 07b] がある。

#### 4. お わ り に

以上、本稿では転移学習を体系的に捉える試みを行い、いろいろな転移学習手法を紹介した。現在の転移学習の研究では、ベクトルを入力とするような識別・回帰問題に対する手法から、木・時系列・グラフなどより複雑な入力を扱うようになってきている [TM 09, TSL 09]。さらに、ドメイン間の関連性の多様化、目標データ数が少ない場合、パラメータ調整のノウハウなども重要だと考えている。

#### 謝 辞

麻生英樹氏、市瀬龍太郎先生、佐久間淳先生、杉山将先生、数原良彦氏には本稿に関する貴重なコメントをいただいた。以上の方々に感謝する。

#### ◇ 参 考 文 献 ◇

- [Ando 05] Ando, R. K. and Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data, *J. Machine Learning Research*, Vol. 6, pp. 1817-1853 (2005)
- [Ando 08] Ando, S. and Suzuki, E.: Unsupervised cross-domain learning by interaction information co-clustering, *Proc. 8th IEEE Int. Conf. on Data Mining*, pp. 13-22 (2008)
- [Argyriou 07] Argyriou, A., Evgeniou, T. and Pontil, M.: Multi-task feature learning, *Advances in Neural Information Processing Systems*, Vol. 19, pp. 41-48 (2007)
- [Argyriou 08] Argyriou, A., Maurer, A. and Pontil, M.: An algorithm for transfer learning in a heterogeneous environment, *Proc. ECML/PKDD2008*, Part I, pp. 71-85 (2008), [LNAI 5211]
- [麻生 03] 麻生英樹, 津田宏治, 村田 昇: パターン認識と学習の統計学—新しい概念と手法, 統計科学のフロンティア, 第 6 巻, 岩波書店 (2003)
- [麻生 09] 麻生英樹: Domain Adaptation とその周辺, 日本音響学会 2009 年春季研究発表会講演論文集, pp. 69-72 (2009)
- [Ben-David 07] Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F.: Analysis of representations for domain adaptation, *Advances in Neural Information Processing Systems*, Vol. 19, pp. 137-144 (2007)
- [Bengio 09] Bengio, Y., Louradour, J., Collobert, R. and Weston, J.: Curriculum learning, *Proc. 26th Int. Conf. on Machine Learning*, pp. 41-48 (2009)
- [Bishop 08] Bishop, C. M., 監訳: 元田 浩ほか, 翻訳: 神島敏弘ほか: パターン認識と機械学習上—ベイズ理論による統計的予測, シュプリンガー・ジャパン (2007-2008)
- [Blitzer 08] Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. and Wortman, J.: Learning bounds for domain adaptation, *Advances in Neural Information Processing Systems*, Vol. 20, pp. 129-136 (2008)
- [Breiman 96] Breiman, L.: Bagging predictors, *Machine Learning*, Vol. 24, pp. 123-140 (1996)
- [Caruana 96] Caruana, R., Baluja, S. and Mitchell, T.: Using the future to “Sort Out” the present: Rankprop and multitask learning for medical risk evaluation, *Advances in Neural Information Processing Systems*, Vol. 8, pp. 959-965 (1996)
- [Caruana 97] Caruana, R.: Multitask learning, *Machine Learning*, Vol. 28, pp. 41-75 (1997)
- [Chapelle 06] Chapelle, O., Schölkopf, B. and Zien, A., eds.: *Semisupervised Learning*, MIT Press (2006)
- [Crammer 08] Crammer, K., Kearns, M. and Wortman, J.: Learning from multiple sources, *J. Machine Learning Research*, Vol. 9, pp. 1757-1774 (2008)
- [Dai 07a] Dai, W., Xue, G.-R., Yang, Q. and Yu, Y.: Coclustering based classification for out-of-domain documents, *Proc. 13th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 210-219 (2007)
- [Dai 07b] Dai, W., Yang, Q., Xue, G.-R. and Yu, Y.: Boosting for transfer learning, *Proc. 24th Int. Conf. on Machine Learning*, pp. 193-200 (2007)
- [Daumé III] Daumé, H., III: Natural language processing blog, <http://nlpers.blogspot.com/search/label/domain%20adaptation>
- [Daumé 06] Daumé, H., III and Marcu, D.: Domain adaptation for statistical classifiers, *J. Artificial Intelligence Research*, Vol. 26, pp. 101-126 (2006)
- [Daumé 07] Daumé, H., III: Frustratingly easy domain adaptation, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256-263 (2007)
- [Do 06] Do, C. B. and Ng, A. Y.: Transfer learning for text classification, *Advances in Neural Information Processing Systems*, Vol. 18, pp. 299-306 (2006)
- [Eaton 08] Eaton, E., desJardins, M. and Lane, T.: Modeling transfer relationships between learning tasks for improved inductive transfer, *Proc. ECML/PKDD2008*, Part I, pp. 317-332 (2008), [LNAI 5211]
- [Freund 96] Freund, Y. and Schapire, R. E.: Experiments with a New Boosting Algorithm, *Proc. 13th Int. Conf. on Machine Learning*, pp. 148-156 (1996)
- [フロイド 99] フロイド, Y., シャピリ, R., 阿部直樹: ブースティング入門, 人工知能学会誌, Vol. 14, No. 5, pp. 771-780 (1999)
- [Freund 03] Freund, Y., Iyer, R., Schapire, R. E. and Singer, Y.: An efficient boosting algorithm for combining preferences, *J. Machine Learning Research*, Vol. 4, pp. 933-969 (2003)
- [Gao 08] Gao, J., Fan, W., Jiang, J. and Han, J.: Knowledge transfer via multiple model local structure mapping, *Proc. 14th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 283-291 (2008)
- [Heckman 79] Heckman, J.: Sample selection bias as a specification error, *Econometrica*, Vol. 47, pp. 153-161 (1979)
- [Hofmann99] Hofmann, T.: Probabilistic latent semantic analysis, *Uncertainty in Artificial Intelligence*, Vol. 15, pp. 289-296 (1999)
- [星野 09] 星野崇宏: 調査観察データの統計科学—因果推論・選択



- バイアス・データ融合, 確率と情報の科学, 岩波書店 (2009)
- [Huang 07] Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M. and Schölkopf, B.: Correcting sample selection bias by unlabeled data, *Advances in Neural Information Processing Systems*, Vol. 19, pp. 601-608 (2007)
- [Kamishima 09] Kamishima, T., Hamasaki, M. and Akaho, S.: TrBagg: A simple transfer learning method and its application to personalization in Collaborative Tagging, *Proc. 9th IEEE Int. Conf. on Data Mining*, pp. 219-228 (2009)
- [金森 09] 金森敬文, 竹之内高志, 村田 昇: パターン認識, R で学ぶデータサイエンス, 第 5 巻, 共立出版 (2009)
- [Liao 05] Liao, X., Xue, Y. and Carin, L.: Logistic regression with an auxiliary data source, *Proc. 22nd Int. Conf. on Machine Learning*, pp. 505-512 (2005)
- [Ling 08] Ling, X., Dai, W., Xue, G.-R., Yang, Q. and Yu, Y.: Spectral domain-transfer learning, *Proc. 14th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 488-496 (2008)
- [LTL 95] Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems, [http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95\\_LTL/transfer.workshop.1995.html](http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html) (1995)
- [元田 06] 元田 浩, 津本周作, 山口高平, 沼尾正行: データマイニングの基礎, オーム社 (2006)
- [Munro 97] Munro, P. W. and Parmanto, B.: Competition among networks improves committee performance, *Advances in Neural Information Processing Systems*, Vol. 9, pp. 592-598 (1997)
- [中村 97] 中村篤祥: 特集『能動学習』, 情報処理, Vol. 38, No. 7, pp. 557-588 (1997)
- [Pan 08a] Pan, S. J., Kwok, J. T. and Yang, Q.: Transfer learning via dimensionality reduction, *Proc. 23rd National Conf. on Artificial Intelligence*, pp. 677-682 (2008)
- [Pan 08b] Pan, S. J. and Yang, Q.: A survey on transfer learning, technical report HKUST-CS08-08, Dept. of Computer Science and Engineering, Hong Kong Univ. of Science and Technology (2008)
- [Raina 06] Raina, R., Ng, A. Y. and Koller, D.: Constructing informative priors using transfer learning, *Proc. 23rd Int. Conf. on Machine Learning*, pp. 713-720 (2006)
- [Raina 07] Raina, R., Battle, A., Lee, H., Packer, B. and Ng, A. Y.: Self-taught learning: Transfer learning from unlabeled data, *Proc. 24th Int. Conf. on Machine Learning*, pp. 759-766 (2007)
- [Rosenstein 05] Rosenstein, M. T., Marx, Z., Kaelbling, L. P. and Dietterich, T. G.: To transfer or not to transfer, *NIPS-2005 Workshop on Inductive Transfer: 10 Years Later* (2005)
- [Rückert 08] Rückert, U. and Kramer, S.: Kernel-based inductive transfer, *Proc. ECML/PKDD2008, Part II*, pp. 220-233 (2008), [LNAI 5212]
- [Satpal 07] Satpal, S. and Sarawagi, S.: Domain adaptation of conditional probability models via feature subsetting, *Proc. 11th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp. 224-235 (2007), [LNAI 4702]
- [Shi 00] Shi, J. and Malik, J.: Normalized cuts and image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888-905 (2000)
- [Shi 08] Shi, X., Fan, W. and Ren, J.: Actively transfer domain knowledge, *Proc. ECML/PKDD2008, Part II*, pp. 342-357 (2008), [LNAI 5212]
- [Shimodaira 00] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood Function, *J. Statistical Planning and Inference*, Vol. 90, pp. 227-244 (2000)
- [杉山 06] 杉山 将: 共変量シフト下での教師付き学習, 日本神経回路学会誌, Vol. 13, No. 3, pp. 111-118 (2006)
- [杉山 07a] 杉山 将: 非定常環境下での教師付き学習: データの入力分布が変化する場合, 画像ラボ, Vol. 18, No. 10, pp. 1-6 (2007)
- [Sugiyama 07b] Sugiyama, M., Krauledat, M. and Müller, K. R.: Covariate Shift Adaptation by Importance Weighted Cross Validation, *J. Machine Learning Research*, Vol. 8, pp. 985-1005 (2007)
- [数原 09] 数原良彦, 植松幸生, 藤野昭典, 片岡良治: 複数情報源を用いた転移学習によるランキング学習, *WdbDB Forum 2009* (2009)
- [Thrun 96] Thrun, S.: Is learning the n-th thing any easier than learning the first?, *Advances in Neural Information Processing Systems*, Vol. 8, pp. 640-646 (1996)
- [Tishby 99] Tishby, N., Pereira, F. C. and Bialek, W.: The Information Bottleneck Method, *Proc. 37th Annual Allerton Conference on Communications, Control and Computing* (1999)
- [TM 09] ICDM 2009 Workshop: Int. Workshop on Transfer Mining, <http://www.cse.ust.hk/?sinnopan/cfp/icdm09wtm.html> (2009)
- [TSL 09] NIPS 2009 Workshop: Transfer Learning for Structured Data, <http://www.cse.ust.hk/?sinnopan/nips09tld/> (2009)
- [TT 05] NIPS 2005 Workshop — Inductive Transfer: 10 Years Later, <http://iitrl.acadiau.ca/itws05/> (2005)
- [Vilalta 02] Vilalta, R. and Drissi, Y.: A perspective view and survey of meta-learning, *Artificial Intelligence Review*, Vol. 18, pp. 77-95 (2002)
- [Wagstaff 01] Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S.: Constrained K-means clustering with background knowledge, *Proc. 18th Int. Conf. on Machine Learning*, pp. 577-584 (2001)
- [Wang 08] Wang, Z., Song, Y. and Zhang, C.: Transferred dimensionality reduction, *Proc. ECML/PKDD2008, Part II*, pp. 550-565 (2008), [LNAI 5212]
- [渡辺 05] 渡辺澄夫, 萩原克幸, 赤穂昭太郎, 本村陽一, 福水健次, 岡田真人, 青柳美輝: 学習システムの理論と実現, 森北出版 (2005)
- [Wolpert 97] Wolpert, D. H. and Macready, W. G.: No free lunch theorems for optimization, *IEEE Trans. on Evolutionary Computation*, Vol. 1, pp. 67-82 (1997)
- [Wu 04] Wu, P. and Dietterich, T. G.: Improving SVM Accuracy by training on auxiliary data sources, *Proc. 21st Int. Conf. on Machine Learning*, pp. 871-878 (2004)
- [Xing 07] Xing, D., Dai, W., Xue, G.-R. and Yu, Y.: Bridged refinement for transfer learning, *Proc. 11th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp. 324-335 (2007), [LNAI 4702]
- [Xue 08] Xue, G.-R., Dai, W., Yang, Q. and Yu, Y.: Topic-bridged PLSA for Cross-domain text classification, *Proc. 31st Annual ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 627-634 (2008)
- [Zadrozny 04] Zadrozny, B.: Learning and evaluating classifiers under sample selection bias, *Proc. 21st Int. Conf. on Machine Learning*, pp. 903-910 (2004)

2010 年 5 月 20 日 受理

## 著者紹介



神嵐 敏弘 (正会員)

1968 年生。1992 年京都大学情報工学科卒業。1994 年同大学院修士課程修了。同年電子技術総合研究所入所。2001 年博士 (情報学)。同年電子技術総合研究所は産業技術総合研究所へ再編。機械学習とその応用の研究に従事。AAAI, ACM, 電子情報通信学会各会員。