情報処理学会論文誌 Vol.57 No.5 1499–1513 (May 2016)

属性ごとの観測確率を考慮したゼロショット学習

鈴木 雅大[1,a)] 佐藤 晴彦[2] 小山 聡[2] 栗原 正仁[2] 松尾 豊[1]

受付日 2015年8月7日, 採録日 2016年2月8日

概要：ゼロショット学習は，1度も学習したことのないカテゴリの画像を，補助情報を頼りに分類する手法である．ゼロショット学習を実現する様々な手法の中でも，補助情報に属性を用いた属性ベースゼロショット学習が最もよく知られている．しかし既存研究では，各属性の画像特徴量への現れやすさを考慮していなかった．本稿ではこのような度合いを属性ごとの観測確率と呼び，観測確率を含めた新たなモデルを提案した．そして提案したモデルの妥当性の検証および既存研究との比較実験によって，提案手法が既存手法と比較して有効性の高いモデルであることを示す．

キーワード：ゼロショット学習，属性，クラス分類

Zero-shot Learning Based on the Observation Probability of Attributes

Masahiro Suzuki[1,a)] Haruhiko Sato[2] Satoshi Oyama[2]
Masahito Kurihara[2] Yutaka Matsuo[1]

Received: August 7, 2015, Accepted: February 8, 2016

Abstract: Zero-shot learning is a method to classify images of categories which have never been trained with a help of side information. Among various methods, attribute-based zero-shot learning which uses attributes as side information is well known. However, the existing method does not consider the frequency which each attribute appears in image features. In this paper, we called this frequency the observation probability of attributes and proposed a new model with the observation probability. Moreover, we showed that our proposed method was more effective than the existing method by inspection of the validity of the proposed model and comparative experiments with the existing method.

Keywords: zero-shot learning, attributes, classification

1. はじめに

今日，インターネットの普及やコンピュータの処理能力の向上によって，大規模なデータを用いた機械学習がさかんに行われている．このような機械学習では，データを表現する特徴ベクトルとデータのクラスを表すラベルをペアとした訓練事例集合から学習する教師あり学習が知られており，大規模なデータを訓練事例集合として学習することで，ラベルが未知のテスト事例集合に対してより良い推定結果が期待できる．しかし，一般的に訓練事例集合のラベルに含まれないクラスを推定することはできない．よって，たとえ大規模なデータがあっても，あらかじめ想定しうるすべてのクラスのラベルを含んだ訓練事例集合を用意することは困難である．特に画像からの一般物体認識問題の場合は，人間が識別する物体カテゴリは30,000種類あるとされていることから [1]，想定しうるカテゴリの画像を十分収集することはほぼ不可能である．このような問題に対処する方法として，他の類似した知識を利用することで精度を向上させる転移学習 [2] などが注目されている．その中でも画像認識の領域では，分類したいテスト事例のカテゴリが訓練事例集合のラベルにまったく含まれていない場合に，共通する情報を頼りにカテゴリを推定するゼロショット学習（zero-shot learning）[3], [4], [5], [6], [7], [8], [9], [10], [11], [12] がさか

[1] 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8654, Japan
[2] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060–0808, Japan
[a)] masa@weblab.t.u-tokyo.ac.jp

んに研究されている．

ゼロショット学習は画像認識領域における問題設定で，Larochelle らによる研究 [3] 以来様々な手法が提案されている．本稿では画像認識に限定して議論を進めるので．本稿で言及するクラスは，一般物体認識問題におけるカテゴリに該当する．ゼロショット学習において，テスト事例集合のクラスラベルを推定するために必要な “共通する情報” は，補助情報（side information）と呼ばれており，何を補助情報とするかがきわめて重要となる．Lampert らはゼロショット学習の補助情報にセマンティックな属性（attribute）を用いた効果的な属性ベースのゼロショット学習として，Direct Attribute Prediction（DAP）モデルを提案した [4], [10]．

属性ベースのゼロショット学習において，画像の特徴ベクトル（以降画像特徴量とする）への現れやすさは，属性によって異なると考えられる．たとえば，色情報で表された画像特徴量では，色に関する属性（“blue” など）は現れやすいが，色とは関係ない属性（“hunter” など）は画像特徴量に現れにくい．特に現れにくい属性は，画像特徴量から適切に学習できないため，目的であるテスト事例集合のクラス推定に負の影響を与える恐れがある．しかし，このような問題について，DAP モデルをはじめ様々な属性を補助情報としたゼロショット学習 [5], [6], [7], [11], [12] では考慮されていなかった．本研究ではこの現れやすさの度合いを属性ごとの観測確率と呼ぶこととし，観測確率を考慮した新たな属性ベースゼロショット学習のモデルを提案する．そして実験によって提案モデルの妥当性と有効性を評価する．

本稿の貢献は次のとおりである．本稿では，観測確率という各属性の画像特徴量への現れやすさの度合いに着目し，その度合いが属性ベースのゼロショット学習において重要であることを示す．また，観測確率を取り入れた新しい属性ベースゼロショット学習のモデルを提案し，DAP モデル [4], [10] と比較して高い正解率となることを示す．そして，その他のゼロショット学習 [4], [10], [11], [12], [13] とも比較をし，これらの研究のように属性以外の補助情報を追加せずに，同等以上の精度となることを示す．さらに，提案手法が DAP モデルの利点を保持しつつ，DAP モデルでの問題点を解決し，扱えなかった問題設定にも適用可能であることを示す．

本稿の構成は以下のとおりである．2 章で関連研究を示し，3 章で提案手法を提示する．そして 4 章で提案手法の仮定などの検証をした後，5 章で既存手法との比較実験をして，考察する．最後に 6 章で，まとめと今後の展望について述べる．

2. 関連研究

本研究のベースとなる既存研究は Lampert らの DAP モデル [4], [10] である．DAP モデルでは，あらかじめ属性のリストを考え，訓練事例集合とテスト事例集合で考えているすべてのクラスについて属性リストとの関係を定義する．ここでの属性とは，人間がクラスに対して持つか持たないかを判別できるセマンティックな概念であり，**図 1** はその定義の例である．図 1 では “polar bear” や “zebra” といった動物のカテゴリがクラス，“black” や “water” といった概念が属性となっている．DAP モデルはこのように定義した属性を補助情報として用いる，属性ベースゼロショット学習の手法である．

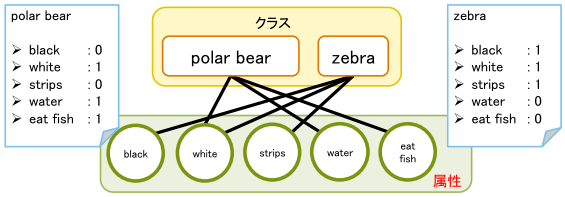

図 1 クラスと属性の例
Fig. 1 Example of classes and attributes.

本研究の提案モデルが DAP モデルと最も異なる点は，観測確率という新しい概念を導入したことである．また，DAP モデルが識別モデルであるのに対して，提案モデルは画像特徴量の生成過程を記述した生成モデルで構築されており，モデルの種類が異なる．さらに，DAP モデルがゼロショット学習しか対応していないのに対して，提案モデルは観測確率によってゼロショット以外の任意のショット数でも学習が可能である．その一方で，DAP モデルの学習に任意の分類器を利用できるという利点は，本研究でも共通である．

本研究で導入する各属性の観測確率は，各属性の画像特徴量への現れやすさの度合いを定式化したものであり，属性ベースゼロショット学習において，本研究で新たに着目した概念である．観測確率と似たような概念に着目した研究として Parikh らの研究がある [14]．Parikh らは，属性には人間にとって理解しやすい（understandable）もしくは識別しやすい（discriminative）属性があり，そのような属性を人間とインタラクションして生成するシステムを提案している．また，Yu らは一般的に人間が画像から認識できるかに応じて属性を visual 属性と non-visual 属性に分けている [5]．このように，観測確率と似た概念はこれまでも着目されてきたが，本質的な部分でこれらの概念と観測確率は異なる．これまでの研究では，通常の画像に対して人間が判別できるかどうかに着目していたが，本研究の観測確率は，画像特徴量に対してモデルもしくは分類器が観測できるかどうかに焦点を当てている．これは，実際に学習や推定を行うのは分類器であり，人間にとって観測できるような属性を選んだところで，必ずしもうまく学習器に認識されるとは限らないからである．このような人間と分類器の判断が異なる現象が発生することは，後述する実

験で実証する．さらに，これまでの研究では学習前に人間が属性を選別する必要があったが，本研究の提案モデルでは，観測確率はパラメータとして含まれているため，学習と同時に観測確率を求め，あらかじめ属性を選別しなくても影響を考慮することができる．

他に観測確率に近い概念として，筆者らは各属性のラベルの偏りや正解率を表す予測能力を定義し，それを DAP モデルに重み付けすることを提案した [15]．しかし，この手法の有効性は DAP モデルと正解率で比較したときわずかな向上しか確認できなかった．また文献 [16] では本稿と同様の観測確率を含めたモデルを提案したが，訓練事例集合とテスト事例集合のそれぞれで求めた観測確率が異なる場合を考慮できていなかった．本稿ではこの点を改善し，さらに複数の検証実験によって，提案手法の有効性を確認している．

生成モデルによる属性ベースゼロショット学習は Yu らにも提案されている [5]．Yu らは Author-Topic モデルをヒントに Category-Topic モデルを提案している．このモデルと本稿の提案モデルが異なる点は，Yu らのモデルがトピックモデルのため画像特徴量として visual word を使わなければならないのに対し，本研究の提案手法は画像特徴量に関する制限がないというところである．そのほかにも，本稿の提案モデルには観測確率が含まれているなど，同じ生成モデルでも異なる点が多い．

Fu らは，クラスによって属性の分布が異なるという問題があることを，projection domain shift 問題として指摘している [12]．Fu らは属性の他に Wikipedia から抽出した言語ベクトル，さらに画像特徴量の 3 つを併用し正準相関分析をすることで，この問題を解決している．本稿の提案手法では，観測確率の事前知識としてテスト事例集合での属性の周辺確率を考慮することで，属性以外の知識を加えずに，訓練事例集合とテスト事例集合での観測確率の違いの解消を試みている．

3. 提案手法

本章では提案手法の内容について述べる．まず 3.1 節で属性ベースのゼロショット学習の問題設定を確認する．3.2 節では，本研究で新たに着目した属性ごとの観測確率について，および観測確率を含めた提案モデルについて説明する．そして 3.3 節で，提案モデルによる学習，推定方法について説明する．

3.1 問題設定

本節では，属性ベースゼロショット学習の問題設定を定式化し，解くべき問題を確認する．

画像の実現値である画像特徴量を $\mathbf{x}$ とし，定義域を $\mathcal{X} = \mathbb{R}^M$ とする．また画像特徴量 $\mathbf{x}$ に対応するクラスを表すクラスラベルを y とし，定義域 $\mathcal{Y}$ を $\{l_1, \ldots, l_K\}$ という K 個のクラス集合とする．

また，本稿ではタスクという概念を用いる．タスクとは，転移学習の枠組みで用いられる用語で，学習や推定するクラスラベルの定義域 $\mathcal{Y}$ を考えたとき，それが同じならば同じタスク，異なるならば異なるタスクとする．ゼロショット学習の場合，訓練事例集合とテスト事例集合はクラスラベルの定義域が異なるため，それぞれを別々のタスクと考える．また，訓練事例集合のクラスラベルの定義域に関して学習や推定を行う問題設定を元タスク，テスト事例集合のクラスラベルの定義域に関して学習や推定を行う問題設定を目標タスクと呼んで区別する．

ある事例が元タスクもしくは目標タスクの事例であることを，それぞれ添え字の s と t で表す．よって元タスクの訓練事例集合は $\mathcal{D}_s = \{\mathbf{X}_s, \mathbf{y}_s\} = \{\mathbf{x}_{si}, y_{si}\}_{i=1}^{N_s}$，目標タスクのテスト事例集合は $\mathcal{D}_t = \{\mathbf{X}_t\} = \{\mathbf{x}_{ti}\}_{i=1}^{N_t}$ と表記する．また，元タスクのクラスラベルの定義域は $\mathcal{Y}_s = \{l_{s1}, \ldots, l_{sK_t}\}$，目標タスクのクラスラベルの定義域は $\mathcal{Y}_t = \{l_{t1}, \ldots, l_{tK_t}\}$ となる．なお本稿の問題設定では，画像特徴量の定義域はどちらのタスクでも同じ M 次元のベクトル空間とする．

ゼロショット学習の目的は，重複しない元タスクと目標タスクのクラス集合（$\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$）が与えられたとき，元タスクの事例集合 $\mathcal{D}_s = \{\mathbf{x}_{si}, y_{si}\}_{i=1}^{N_s}$ を用いて，目標タスクの事例集合 $\mathcal{D}_t = \{\mathbf{x}_{ti}\}_{i=1}^{N_t}$ のクラスラベル $\{y_{ti}\}_{i=1}^{N_t}$ を推定することである．このとき $\mathcal{D}_s$ は訓練事例集合，$\mathcal{D}_t$ はテスト事例集合としてのみ用いることに留意されたい．

属性ベースのゼロショット学習では，補助情報として属性を導入する．属性は M 個与えられていて，その実現値を $\mathbf{a} = [a_1, \ldots, a_M]$ とし，定義域は $\mathcal{A} = \{0, 1\}^M$ とする．以降，属性の実現値のことを，単に属性値と呼ぶ．m 番目の属性値 a_m は，0 ならば m 番目の属性を持たず，1 ならば m 番目の属性を持つことを意味する．

属性値は，事例ごと，すなわち各事例の画像特徴量とクラスラベルの両方 $\{\mathbf{x}_i, y_i\}_{i=1}^{N}$ によって $\mathbf{A}^y = [\mathbf{a}^{y_1}, \ldots, \mathbf{a}^{y_N}]^{\mathrm{T}}$ として定義される．元タスクの場合，属性値は $\mathbf{A}^{y_s} = [\mathbf{a}^{y_{s1}}, \ldots, \mathbf{a}^{y_{sN_s}}]^{\mathrm{T}}$ となる．ただし，クラスラベルが同じになるような事例の属性値をすべて同じにする場合（すなわち $y_i = y_j$ のとき $\mathbf{a}^{y_i} = \mathbf{a}^{y_j}$）は，特にクラスラベルごとの定義と呼ぶ．目標タスクの場合も $\mathbf{A}^{y_t} = [\mathbf{a}^{y_{t1}}, \ldots, \mathbf{a}^{y_{tN_t}}]^{\mathrm{T}}$ のように事例ごとに属性値が定義されるが，実際の問題設定では事例集合が与えられないため，クラス集合の任意のクラス l_t に対する定義 $\mathbf{A}^{l_t} = [\mathbf{a}^{l_{t1}}, \ldots, \mathbf{a}^{l_{tK_t}}]^{\mathrm{T}}$ を事前情報としてクラス推定に利用する．

以上の議論を**表 1** にまとめる．

3.2 観測確率を考慮した属性ベースゼロショット学習

本節では，属性ベースゼロショット学習を解くための新たな手法として，観測確率という概念と，それを考慮した属性ベースゼロショット学習の提案モデルの説明をする．

表 1 属性ベースゼロショット学習の問題定式化
Table 1 Problem formulation of attribute-based zero-shot learning.

事前情報	$\mathcal{Y}_s$，$\mathcal{Y}_t$（ただし $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$） $\mathbf{A}^{y_s}$，$\mathbf{A}^{l_t}$
訓練事例集合	$\mathcal{D}_s = \{\mathbf{x}_{si}, y_{si}\}_{i=1}^{N_s}$
テスト事例集合	$\mathcal{D}_t = \{\mathbf{x}_{ti}\}_{i=1}^{N_t}$
学習目的	テスト事例集合のクラスラベルを推定する

図 2 属性による画像特徴量への現れやすさの違い
Fig. 2 Difference in the frequency which each attribute appears in image features.

3.2.1 観測確率

それぞれの属性が入力である画像特徴量にどれくらい現れるか，すなわち画像特徴量に対する各属性の現れやすさについては，属性によって異なると考えられる．たとえば，クジラの画像が RGB color histogram のような色の画像特徴量で表されているとする（**図 2**）．“blue” という属性は色の知識なので画像特徴量に現れやすいが，“hunter” 属性は色情報では解釈しにくいので，この種類の画像特徴量には現れにくい．このように，用いる画像特徴量の種類が同じでも，属性が異なれば画像特徴量への現れ方は異なると考えられる．

この現れやすさは，各属性の属性値 a_m を画像特徴量 $\mathbf{x}$ から任意の分類器 f_m で学習するときに大きく影響する．特に現れにくい属性の場合，画像特徴量にその属性に該当するような情報が少ないため，分類器 f_m は適切に学習できないと考えられる．そのため，分類器 f_m から推定結果として得られる確信度 $p(a_m|\mathbf{x})$ も，同様に正しく求められない．属性ベースゼロショット学習の代表的な手法である DAP モデル [4], [10] では，クラスラベル y の事後確率 $p(y|\mathbf{x})$ を確信度 $p(a_m|\mathbf{x})$ を用いて $p(y|\mathbf{x}) \propto \prod_m \frac{p(a_m^y|\mathbf{x})}{p(a_m^y)}$ として求めて，クラスラベル y の推定をしている．しかし，画像特徴量への現れやすさを考慮していないため，ある属性 a_m が画像特徴量 $\mathbf{x}$ に現れにくい場合，確信度 $p(a_m|\mathbf{x})$ が正しく推定できなくなり，その影響からクラスラベル y の事後確率 $p(y|\mathbf{x})$ の良い推定を得られなくなる恐れがある．

以上の理由から，本研究では画像特徴量に対する各属性の現れやすさという度合いが重要であると考え，この度合いを考慮した手法を提案する．今後の議論のためにこの度合いを属性ごとの観測確率と呼ぶこととする．

3.2.2 提案モデル

本研究では，観測確率を導入した新たな属性ベースゼロショット学習の生成モデルを提案する．生成モデルとは，データの生成過程を明示的に記述するモデルであり，本研究ではクラスラベルから画像特徴量が生成される過程をモデル化する．**図 3** は提案する生成モデルをグラフィカルモデルで表したものである．グラフィカルモデルとは，生成モデルで設計した確率変数の依存関係を有向グラフで表現したもので，観測変数[*1]を黒丸，潜在変数を白丸で図示している．この生成過程は，元タスク・目標タスク両方で共通とする．

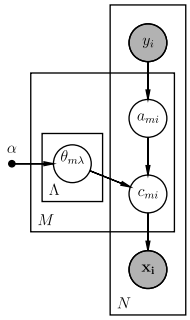

図 3 提案モデルのグラフィカルモデル
Fig. 3 Graphical model of our proposed model.

3.2.1 項の例で，画像特徴量から観測される属性値と，クラスラベルに対してあらかじめ定義されている属性値が異なる場合があることを示した．本研究ではこの 2 つを異なる属性として明確に区別し，前者を観測された属性と呼び，その実現値を $\mathbf{c} = [c_1, \ldots, c_M]$，定義域を $\mathcal{C} = \{0,1\}^M$ とする．また後者については，3.1 節で設定した属性と同義であるため，その属性値も同様に $\mathbf{a}$ と表記する．ただし観測された属性と区別する場合は，真の属性と呼ぶ．この 2 つの属性値が同じならば，その属性は画像特徴量に現れやすく，異なるならば現れにくいということになる．

次に，生成過程とその分布の形について順番に説明する．事例 i のクラスラベル y_i が与えられれば，事前情報として与えられている定義 $\mathbf{a}^{y_i}$ によって真の属性の属性値が定まる．ただし，属性値の定義は事例ごとにされているため，クラスラベルが同じでも真の属性値が同じ値とは限らないことに留意する．この生成過程は $\mathbf{a}_i \sim p(\mathbf{a}_i|y_i) = [[\mathbf{a}_i = \mathbf{a}^{y_i}]]$ と表現できる．ただし $[[P]]$ は Iverson の記法で，P が真のとき 1，偽のとき 0 をとる．

真の属性から観測された属性への生成過程を考えると，生成される間に何からの要因によって，属性値が変化していると考えられ，その変化の確率は現れやすさ，すなわち観測確率と同義である．よって観測確率を，m 番目の真の属性の属性値 $\lambda = a_{mi}$ から，観測された属性の属性値 c_{mi} を生成する確率変数 $\theta_{m\lambda}$ として定義する．属性値の定義域は $\{0,1\}$ の 2 値なので，観測された属性はベルヌーイ分布に従って生成されると考え，観測確率はベルヌー

[*1] ここでの観測変数は，実際に観測値として与えられるモデル内の確率変数を指し，本稿で導入する観測確率とは関係ない．

イ分布のパラメータとする．したがって，観測された属性の生成過程は $c_{mi} \sim p(c_{mi}|a_{mi}, \theta_{m\lambda}) = p(c_{mi}|\theta_{ma_{mi}}) = Bern(c_{mi}|\theta_{ma_{mi}}) = \theta_{ma_{mi}}^{c_{mi}}(1-\theta_{ma_{mi}})^{1-c_{mi}}$ となる．

さらに，ベルヌーイ分布の共役事前分布はベータ分布なので，パラメータである観測確率 $\theta_{m\lambda}$ はハイパーパラメータ $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]$ によって $\theta_{m\lambda} \sim p(\theta_{m\lambda}; \boldsymbol{\alpha}) = Beta(\theta_{m\lambda}|\boldsymbol{\alpha}) = \frac{\theta_{m\lambda}^{\alpha_0-1}(1-\theta_{m\lambda})^{\alpha_1-1}}{B(\boldsymbol{\alpha})}$ のように生成される．ただし，$B(\boldsymbol{\alpha})$ はベータ関数である．

画像特徴量 $\mathbf{x}_i$ は観測された属性から $\mathbf{x}_i \sim p(\mathbf{x}_i|\mathbf{c}_i)$ として生成される．ただし，この生成過程の形については次節で説明する．

この生成モデルでは観測確率について次の 2 つの仮定をしている．

仮定 1 観測確率は属性によって値が異なる．

仮定 2 画像特徴量の種類が同じならば，異なるタスクで観測確率が等しい．

これらは提案モデルの前提となる仮定であり，実験でこれらの仮定の妥当性を検証する必要がある．

3.3 提案モデルの学習・推定

提案手法では，元タスクの訓練事例集合から提案モデルのパラメータを学習する**訓練段階**と，目標タスクで元タスクで学習したパラメータを再利用し，テスト事例集合からクラスラベルを推定する**推定段階**がある．

3.3.1 項と 3.3.2 項では提案モデルから学習段階および推定段階の計算を導出する．3.3.3 項では，観測確率を異なるタスクで近づける工夫を説明する．

本節で最終的に得られる学習段階と推定段階のアルゴリズムをそれぞれ Algorithm 1 と Algorithm 2 に示す．また，**図 4** で学習，推定段階の概要を示す．

3.3.1 学習段階

学習段階では，元タスクの訓練事例集合 $\mathcal{D}_s$ が与えられたときの提案モデルのパラメータ，すなわち観測確率を学習する．

生成モデルからパラメータを推定する一般的な方法として，EM アルゴリズムがよく知られているが，本研究ではそれとは異なるアプローチをとる．これは，既存研究の DAP モデルには，学習段階で画像特徴量に応じて任意の分類器を利用できるという特徴があり，本研究の提案モデルでもその特徴を残すためである．提案モデルでは以降説明する工夫によって，生成モデルでありながら DAP モデルと同様に，任意の分類器で学習できる．またこの工夫によって，真の属性，観測された属性および観測確率に明確な解釈が与えられる．

モデルの設計の段階で $p(\mathbf{x}_i|\mathbf{c}_i)$ の分布の形を明示化しなかったが，この分布をベイズの定理により $p(\mathbf{x}_i|\mathbf{c}_i) = \frac{p(\mathbf{c}_i|\mathbf{x}_i)p(\mathbf{x}_i)}{p(\mathbf{c}_i)} = \prod_m \frac{p(c_{mi}|\mathbf{x}_i)p(\mathbf{x}_i)}{p(c_{mi})}$ とする．ここで観測された属性が独立，および画像特徴量に対して条件付き独立であると仮定している．この仮定は，DAP モデルでも暗黙的に設定されていたものであり，本稿でもこの仮定を採用する．

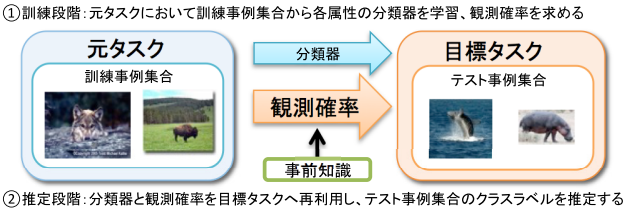

図 4 学習段階と推定段階の概要図

Fig. 4 Outline of the training phase and the estimation phase.

Algorithm 1 Training algorithm of our proposed model

Input: $\mathcal{D}_s$, $\mathbf{A}^{y_s}$
Output: Θ^*, $\mathbf{f}$

Separate $\mathcal{D}_s$ into $\mathcal{D}_{tr}$ and $\mathcal{D}_v$
for m in $\{1, \ldots, M\}$ **do**
  Train probabilistic classifier $f_m : \mathcal{X} \to \mathcal{A}_m$
  $\theta^*_{m0} = \theta^*_{m1} = N_{vm0} = N_{vm1} = 0$
  for i in $\{1, \ldots, N_v\}$ **do**
    Estimate $p(c_{mi}|\mathbf{x}_{vi})$ from f_m
    if $a_m^{y_i} == 0$ **then**
      $\theta^*_{m0} = \theta^*_{m0} + p(c_m = 1|\mathbf{x}_{vi})$
      $N_{vm0} = N_{vm0} + 1$
    else
      $\theta^*_{m1} = \theta^*_{m1} + p(c_m = 1|\mathbf{x}_{vi})$
      $N_{vm1} = N_{vm1} + 1$
    end if
  end for
end for
return $\Theta^* = \{\theta^*_{m0}, \theta^*_{m1}, N_{vm0}, N_{vm1}\}_{m=1}^M$, $\mathbf{f} = \{f_m\}_{m=1}^M$

Algorithm 2 Estimation algorithm of our proposed model

Input: $\mathcal{D}_t$, Θ^*, $\mathbf{f}$, $\mathbf{A}^{l_t}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$
Output: $\mathbf{y}_t$

for i in $\{1, \ldots, N_t\}$ **do**
  for m in $\{1, \ldots, M\}$ **do**
    Estimate $p(c_{mi}|\mathbf{x}_{ti})$ from f_m
  end for
end for
for m in $\{1, \ldots, M\}$ **do**
  $p(c_m) = \frac{\sum_i p(c_{mi}|\mathbf{x}_{ti})}{N_t}$
  for λ in $\{0, 1\}$ **do**
    $\alpha_\lambda^{sample} = 0$
    for j in $\{1, \ldots, \eta_\lambda N_{vm\lambda}\}$ **do**
      $Sample \sim p(c_m = \lambda)$
      $\alpha_\lambda^{sample} = \alpha_\lambda^{sample} + Sample$
    end for
  end for
  $\theta_{m0} = \frac{\theta^*_{m0} + \alpha_0 + \alpha_0^{sample} - 1}{N_{vm0} + \alpha_0 + \alpha_0^{sample} + \alpha_1 + \alpha_1^{sample} - 2}$
  $\theta_{m1} = \frac{\theta^*_{m1} + \alpha_0 + \alpha_1^{sample} - 1}{N_{vm1} + \alpha_0 + \alpha_0^{sample} + \alpha_1 + \alpha_1^{sample} - 2}$
end for
for i in $\{1, \ldots, N_t\}$ **do**
  Estimate y_{ti} by Eq. (5)
end for
return $\mathbf{y}_t = \{y_{ti}\}_{i=1}^{N_t}$

次に，条件付き独立とした条件付き確率 $p(c_{mi}|\mathbf{x}_i)$ に着目する．本稿ではこの条件付き確率を，分類器 $f_m : \mathcal{X} \to \mathcal{A}_m$ の確信度として得たものと考える．その理由を以下で説明する．

元タスクの訓練事例集合 $\mathcal{D}_s$ を $\mathcal{D}_{tr}$ と $\mathcal{D}_v$ に分け，$\mathcal{D}_{tr}$ を訓練事例集合として画像特徴量から属性への写像を得る．具体的には各属性の属性値 a_m についてそれぞれ任意の分類器 f_m を用意し $f_m : \mathcal{X} \to \mathcal{A}_m$ を学習する．なお，この学習に必要な各事例の属性値ラベル $\mathbf{a}_i$ はあらかじめ定義されていることに留意されたい．学習後 $\mathcal{D}_v$ を検証用事例集合とし，その画像特徴量 $\mathbf{x}_{vi}$ を各属性の分類器 f_m の入力として与えることで出力として各属性の確信度 $\hat{p}(a_{mi}|\mathbf{x}_{vi})$ を得る．しかし，この確信度は真の属性の確信度ではなく，学習していない未知の検証用事例集合から属性を観測できた度合いである．よって，この確信度を観測された属性の c_{mi} の確信度 $p(c_{mi}|\mathbf{x}_{vi})$ とする．

以上の議論によって，真の属性，観測された属性および観測確率について，分類器 f_m によって明確な解釈が与えられる．真の属性とは，分類器 f_m のラベルとして与えられる値である．また観測された属性は，分類器 f_m が未知の画像特徴量から得られた値であり，その度合いは確信度として得られる．そして観測確率は，分類器 f_m が画像特徴量から属性を観測できた確率であり，生成モデルとしては定義された属性値と未知の画像特徴量から得られた属性値の違いの原因として解釈できる．

$\mathcal{D}_v$ の画像特徴量 $\mathbf{X}_v$ が与えられたときの観測確率の事後確率は $p(\theta_{m\lambda}|\mathbf{X}_v)$ であり，訓練段階では，この確率を最大にするような観測確率 $\theta_{m\lambda}$ を求めたい．このように事後確率を最大化してパラメータを推定する手法を最大事後確率推定（MAP 推定）という．すなわち次の式を解く必要がある．

$$\hat{\theta}_{m\lambda} = \arg\max_{\theta_{m\lambda}} p(\theta_{m\lambda}|\mathbf{X}_v) \tag{1}$$

$p(\theta_{m\lambda}|\mathbf{X}_v)$ を直接最大化することは困難だが，対数をとって下界を求めると解析的に解くことができ，次のように求まる．

$$\hat{\theta}_{m\lambda} = \frac{\sum_{i:a_m^{y_{vi}}=\lambda} p(c_{mi}=1|\mathbf{x}_{vi}) + \alpha_0 - 1}{N_{vm\lambda} + \alpha_0 + \alpha_1 - 2} \tag{2}$$

ただし $p(c_{mi}=1|\mathbf{x}_{vi})$ は検定用事例集合から得た確信度，$N_{vm\lambda} = \sum_{i:a_m^{y_{vi}}}$ である．式 (2) の詳しい導出は付録 A.1 に記載する．

なお，式 (2) の $\sum_{i:a_m^{y_{vi}}}$ の部分から，$\mathbf{x}_{vi}$ にクラスラベル y_{vi} が与えられていないと，式 (2) は解くことができないことに留意する．

3.3.2 推定段階

推定段階では，学習段階で学習した観測確率と分類器を再利用して，目標タスクのテスト事例集合 $\mathcal{D}_t$ のクラスラベルを推定する．

まず，提案モデルからクラスラベル y_{ti} の事後確率 $p(y_{ti}|\mathbf{x}_{ti})$ を求める式を導出する．

テスト事例集合 $\mathcal{D}_t$ の画像特徴量 $\mathbf{x}_{ti}$ およびクラスラベル y_{ti} の同時確率は，提案モデルより次のようになる．

$$\begin{aligned}
&p(\mathbf{x}_{ti}, y_{ti}|\boldsymbol{\theta}) \\
&= \sum_{\mathbf{c}}\sum_{\mathbf{a}} p(\mathbf{x}_{ti}, y_{ti}, \mathbf{c}, \mathbf{a}|\boldsymbol{\theta}) \\
&= p(y_{ti}) \prod_m \sum_{c_m}\sum_{a_m} p(a_m|y_{ti}) p(\mathbf{x}_{ti}|c_m) p(c_m|a_m, \theta_{m\lambda}) \\
&\simeq p(y_{ti}) p(\mathbf{x}_{ti}) \prod_m \sum_{c_m} \frac{p(c_m|\theta_{ma_m^{y_{ti}}}) p(c_m|\mathbf{x}_{ti})}{p(c_m)}
\end{aligned} \tag{3}$$

式 (3) から画像特徴量 $\mathbf{x}_{ti}$ に対するクラスラベル y_{ti} の事後確率は，次のようになる．

$$\begin{aligned}
p(y_{ti}|\mathbf{x}_{ti}) &= \frac{p(\mathbf{x}_{ti}, y_{ti})}{p(\mathbf{x}_{ti})} \\
&= p(y_{ti}) \prod_m \sum_{c_m} \frac{p(c_m|\theta_{ma_m^{y_{ti}}}) p(c_m|\mathbf{x}_{ti})}{p(c_m)} \\
&\propto \prod_m \sum_{c_m} \frac{p(c_m|\theta_{ma_m^{y_{ti}}}) p(c_m|\mathbf{x}_{ti})}{p(c_m)}
\end{aligned} \tag{4}$$

ただし，式変形の中で $p(y_{ti})$ を一様分布と仮定した．

式 (4) を計算するためには，$p(c_m|\theta_{ma_m^y})$，$p(c_m|\mathbf{x})$，$p(c_m)$ をそれぞれ求める必要がある．

確信度 $p(c_m|\mathbf{x})$ は，元タスクの訓練事例集合で学習した分類器 f_m を再利用し $p(c_m|\mathbf{x}) = f_m(\mathbf{x})$ として推定できる．

周辺確率 $p(c_m)$ は $p(c_m) = \int p(\mathbf{x}, c_m) d\mathbf{x} = \int p(\mathbf{x}) p(c_m|\mathbf{x}) d\mathbf{x} \simeq \frac{1}{N_t}\sum_i p(c_{mi}|\mathbf{x}_{ti})$ として求められる．ただし $p(c_{mi}|\mathbf{x}_{ti})$ は分類器 f_m から求めた確信度である．

$p(c_m|\theta_{ma_m^y})$ はベルヌーイ分布であり，分布のパラメータ，すなわち観測確率 $\theta_{m\lambda}$ は，訓練段階において元タスクで推定したものを再利用する．

推定するクラスラベル $\hat{y}_{ti}$ は目標タスクのクラス集合 $\mathcal{Y}_t$ の要素のいずれかであり，式 (4) の事後確率が最大になるようなクラスを選択する（MAP 推定）．

$$\begin{aligned}
\hat{y}_{ti} &= \arg\max_{l_{tk}\in\mathcal{Y}_t} p(y_{ti} = l_{tk}|\mathbf{x}_{ti}) \\
&= \arg\max_{l_{tk}\in\mathcal{Y}_t} \prod_m \sum_{c_m} \frac{p(c_m|\theta_{ma_m^{l_{tk}}}) p(c_m|\mathbf{x}_{ti})}{p(c_m)}
\end{aligned} \tag{5}$$

この推定は，目標タスクのすべてのクラスについて，それぞれ式 (4) の事後確率を計算し，それらの中で最大となる事後確率のクラスを選択することで，容易に求まる．

3.3.3 タスク間の観測確率の違いの補正

ここで仮定 2 の妥当性について，式 (2) で求めた観測確率の推定式から検討する．式 (2) 中の観測された属性の事後確率 $p(c_{mi}=1|\mathbf{x}_i)$ は，画像特徴量の種類が同じならば，

事例が異なっても似た傾向になると考えられる．この前提に従えば，求めた観測確率も事例によらないことになり，仮定 2 は成り立つと考えられる．しかし，属性値は定義より，事例，すなわち画像特徴量とクラスラベルが与えられたときに一意に定まるため，事例が異なる場合には事後確率は厳密には異なる値となる．つまり，元タスクの検定用事例を増やして観測確率の精度を上げようとしても，属性値の定義が異なるため，目標タスクをテスト事例としたときの観測確率とは同じにはならない．よって，目標タスクの推定に利用する観測確率は，正確には同じ目標タスクの事例の事後確率 $p(c_{mi}=1|\mathbf{x}_{ti})$ から求めた観測確率でなければならない．この観測確率は，目標タスクの事例を用いて得られた観測確率であるため，目標タスクの観測確率 θ^{target} と呼ぶ．しかし実際の問題設定では目標タスクのクラスラベルが得られないため，事後確率を求めても式 (2) から目標タスクの観測確率を計算することができない．

この問題を解決するため，本稿ではタスク間の観測確率の違いを補正する手法を提案する．式 (2) の $\boldsymbol{\alpha}$ は，元々観測確率を生成するベータ分布 $Beta(\theta_{m\lambda}|\boldsymbol{\alpha})=\frac{\theta_{m\lambda}^{\alpha_0-1}(1-\theta_{m\lambda})^{\alpha_1-1}}{B(\boldsymbol{\alpha})}$ のハイパーパラメータであり，どれだけ属性値を観測したかという有効観測数に該当する．さらにハイパーパラメータを m 番目の属性ごと（$\alpha_{m\lambda}$）に考えることで，各属性の観測数，すなわち事前知識を取り入れることが可能となる．本手法では，目標タスクの属性値に従ってハイパーパラメータを調節することで，観測確率が目標タスクに近づくように補正する．

目標タスクで求めた事例ごとの事後確率 $p(c_{mi}=1|\mathbf{x}_{ti})$ から，観測された属性の周辺確率 $p(c_m)$ を求める．この周辺確率は目標タスクにおいて属性値 c_m の観測される傾向を表している．これを事前知識として反映させるために，周辺確率 $p(c_m=\lambda)$ をベルヌーイ分布として任意の数サンプリングし，その総和 $\alpha_{m\lambda}^{sample}$ をハイパーパラメータとして式 (2) の分子に加える．また，式 (2) の分母には正規化のため $p(c_m=0)$ と $p(c_m=1)$ のサンプリング総和 α_{m0}^{sample} と α_{m1}^{sample} を加える．よって，タスクの違いを補正した観測確率の推定式は次のようになる．

$$\hat{\theta}_{m\lambda}=\frac{\sum_{i:a_m^{y_{vi}}=\lambda}p(c_{mi}=1|\mathbf{x}_{vi})+\alpha_0+\alpha_{m\lambda}^{sample}-1}{N_{vm\lambda}+\alpha_0+\alpha_{m0}^{sample}+\alpha_1+\alpha_{m1}^{sample}-2} \tag{6}$$

事前知識を取り入れる程度はサンプリング回数によって決まる．本稿では，元タスクの検証用事例集合の総数 $N_{vm\lambda}$ に対する割合 η_λ で調節する．ただし割合 η_λ を大きくしすぎると，観測確率 $\theta_{m\lambda}$ が $p(c_m=\lambda)$ と等価になってしまうので，なるべくサンプリング回数を検定用事例集合数以下（$\eta_\lambda\leq 1$）にするべきと考えられる．

なお，本手法では目標タスクの事後確率を求めないと観測確率が求まらないため，元タスクで正規化前の観測確率 $\theta_{m\lambda}^{*}=\sum_{i:a_m^{y_{vi}}=\lambda}p(c_{mi}=1|\mathbf{x}_{vi})$ と正規化定数 $N_{vm\lambda}=\sum_{i:a_m^{y_{vi}}=\lambda}$ を求めておく．推定段階でこれらを再利用し，目標タスクでのサンプリング値から，式 (6) を計算して観測確率を求める．

学習段階と推定段階のアルゴリズムの擬似コードである Algorithm 1 と Algorithm 2 では，本節で述べたサンプリングの方法を用いた学習・推定アルゴリズムが記述されている．

4. 検証実験

本章では既存研究との比較実験の前に，提案モデルで設定した仮定などについて実験によって検証し，提案手法の妥当性について議論する．

4.2 節で仮定 1 と仮定 2 の検証実験をし，4.3 節で人間と分類器での属性の観測しやすさの違いを検証して観測確率を用いる妥当性について議論する．そして 4.4 節で 3.3.3 項で提案した目標タスクの観測確率に近づける手法の有効性を検証する．

4.1 データセット

本稿では Lampert ら [4], [10] によって作成された Animals with Attributes（以下 AwA）[*2] と Farhadi ら [17] によって作成された aPascal-aYahoo（以下 aP-aY）[*3] の 2 種類のデータセットで実験を行った．これら 2 つのデータセットの最も大きな違いは，aP-aY では属性が事例ごとに定義されている一方で，AwA ではクラスごと（$y_i=y_j$ のとき $\mathbf{a}^{y_i}=\mathbf{a}^{y_j}$）に定義されているという点である．また AwA の属性は画像とは無関係なものが多い一方で，aP-aY は画像から認識しやすいような属性が付けられている．その他の違いは**表 2** にまとめた．

4.2 検証 1：仮定 1 と仮定 2 の検証

本実験では，提案モデルでの 2 つの仮定を検証する．

用いるデータセットは AwA とし，元タスクと目標タスクのクラスは，それぞれデフォルトで分けられている訓練事例集合とテスト事例集合のクラスラベルの定義域と

表 2 データセットの違い
Table 2 Differences in datasets.

データセット	Animals with Attributes	aPascal-aYahoo
画像数	30,475	15,339
クラス数	50	32（Pascal 20・Yahoo 12）
属性の数	85	64
属性の定義	クラスごと	事例ごと
クラスの種類	動物	動物・乗り物・家電

[*2] http://attributes.kyb.tuebingen.mpg.de/
[*3] http://vision.cs.uiuc.edu/attributes/

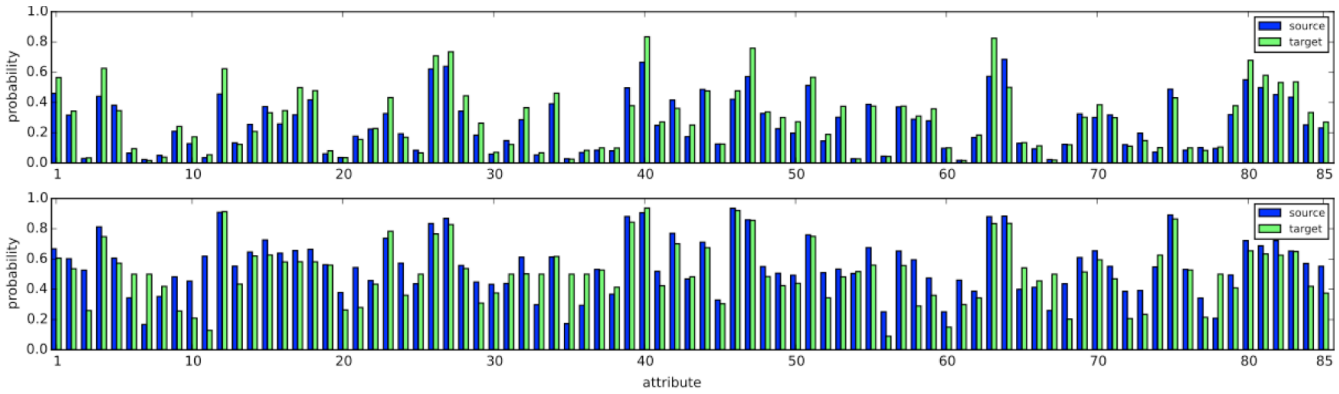

図 5 元タスクの観測確率と目標タスクの観測確率の比較

Fig. 5 Comparison of the observation probabilities on the source task and on the target task.

した（元タスクが 40 クラス，目標タスクが 10 クラス）．元タスクでの検証用事例数は，元タスクの事例集合全体の 10%とした．ハイパーパラメータは $\alpha_0 = \alpha_1 = 2$ とした．なお，本実験では 3.3.3 項で示したサンプリングの方法は利用していない．画像特徴量の種類は 4,096 次元の DeCAF（Deep Convolutional Activation Feature）[18] を利用した．DeCAF は ILSVRC2012 データセットで事前学習済みの 7 層の畳み込みディープニューラルネットワークに，画像を入力として与えたときの中間層の出力を特徴量としたもので，本研究では fc7 層の出力を特徴量とした．分類器には L2 正則化ロジスティック回帰を用いた．検証 1 では，分類器を固定したうえで提案手法による効果を確認するため，ロジスティック回帰のパラメータ C は 1.0 と固定した．実装は Python 2.7 [*4] で行い，機械学習のライブラリ scikit-learn 0.15.2 [*5] を利用した．

仮定 1 と仮定 2 が妥当かどうか検証するために，各属性の元タスクで求めた観測確率と目標タスクの観測確率をプロットする．AwA はクラスごとに用意されている事例数が異なるので，偏りをなくすため各クラスの事例数を 90 枚とした．3.3.3 項でも述べたとおり，目標タスクの観測確率は実際の問題設定では求められないが，この実験では目標タスクにもラベルがあるものとして計算している．

図 5 が検証結果である．上の図が元タスクで求めた観測確率 θ_{m0} と目標タスクの観測確率 θ_{m0}^{target} の値で，下の図が θ_{m1} と θ_{m1}^{target} の値を示す．横軸が m 番目の属性，縦軸が観測確率の値を表す．青が元タスクの観測確率で，緑が目標タスクの観測確率を表す．

まず，元タスクの観測確率について着目する．図 5 から，仮定 1 で仮定したとおり属性によって観測確率の値が大きく異なることが分かる．もしすべての属性が完全に画像に現れて適切に学習できれば，θ_{m0} はすべて 0，θ_{m1} はすべて 1 となっているはずである．図 5 をみると全体の傾向として θ_{m0} は 0 に近く，θ_{m1} は 1 に近いように分布しているが，個々の属性にはかなりばらつきがあることが分かる．

次に図 5 について，観測確率と目標タスクの観測確率の分布の違いに着目する．この図から仮定 2 で仮定したとおり，おおよそ近い分布になっていることが確認できる．しかし完全に一致しているわけではなく，3.3.3 項で議論したようにタスクの違いが影響していると考えられる．

以上の検証によって提案モデルでの仮定はおおむね妥当であることを示した．

4.3 検証 2：観測確率の妥当性の検証

提案手法では，分類器の各属性の確信度から属性ごとの観測確率を求めている．似たような概念として，Parikh ら [14] は人間にとって理解しやすい属性を考え，人間とのインタラクションによってこのような属性を生成するシステムを提案している．また，Yu ら [5] は AwA の属性のうち一般的に人間が画像から理解できるであろう属性を visual 属性とし，それ以外を non-visual 属性として区別した．

本節では，分類器にとって画像特徴量から観測しやすい属性と人間が通常の画像から判断できる属性を比較検証し，本研究の問題設定において，観測確率を用いることの妥当性について議論する．

観測確率が高い属性と低い属性がどのようなものかを検証するために，図 5 の結果から，観測確率による属性の上位 5 個と下位 5 個を**表 3** に示した．表中の太文字の属性が Yu らによる visual 属性である．また，visual 属性と観測確率の関係の定量的な評価値として Area Under the Curve（AUC）による評価も載せている．表中の AUC は観測確率が visual 属性と non-visual 属性を適切に分類できたかを表している．表 3 では，θ_{m1} が順位を適切に分類できた属性の順位であるのに対し，θ_{m0} は逆に順位が低い属性がより適切に分類できた属性となることに注意する．これは AUC による評価値も同様で，θ_{m0} のときは 1 で完全

[*4] https://www.python.org
[*5] http://scikit-learn.org

表 3 観測確率による属性の順位と AUC による評価（太文字の属性は visual 属性）

Table 3 Ranking of attributes according to the observation probabilities and evaluation by AUC (Visual attributes are indicated in bold).

観測確率	上位 5 位	下位 5 位	AUC
θ_{m0}	oldworld fast **chewtheeth** **tail** newworld	skimmer **desert** **red** flys plankton	0.365
θ_{m1}	quadrapedal **furry** fast **ground** oldworld	**red** flys **cave** scavenger insects	0.431

に正しく分類，0.5 でランダム，0 ですべて逆に分類したことになるが，θ_{m1} ではすべて逆となる．

表 3 から，visual 属性，すなわち人間が画像から判断できるような属性が，必ずしも高い観測確率になるとは限らないということが分かる．θ_{m1} の oldworld や fast などは，人間は明らかに画像に現れないと判断すると思われる属性である．さらに，oldworld や fast，red などのように θ_{m0} と θ_{m1} のどちらにも上位または下位にくる属性がよく見られる．これは，属性ラベルがインバランスであるなどの理由で，真の属性の属性値がどちらの場合も観測された属性の片方の属性値しか観測できない状態になっているためと考えられる．また，θ_{m0} と θ_{m1} の両方の場合で AUC が 0.5 に近いことから，どちらの場合もほぼランダムな分類となっていることが分かる．

以上のように，分類器にとって画像特徴量から観測しやすい属性と，人間が通常の画像から判断できる属性には，関係性が低いことが示された．この結果から，あらかじめ人間が分類器が適切に学習できそうな属性を選択することには限界があると考えられる．また，そのような属性を人間が判断しようとすると，それだけ人的コストもかかってしまう．その一方，観測確率を用いれば，人間があらかじめ判別せずに学習段階で分類器が画像特徴量からうまく学習できるかどうかを推定することができる．よって本研究の問題設定では，人手で定義した属性の観測しやすさではなく，観測確率を用いる方が妥当であると考えられる．

4.4 検証 3：“タスク間の観測確率の違いの補正” の検証

3.3.3 項で，観測確率を目標タスクの観測確率に近づける方法として，サンプリングによるタスク間の違いの軽減を提案した．本実験では，この手法の効果を検証実験によって評価する．AwA のすべての訓練事例集合を利用し，検証用事例集合は訓練事例集合のうち各クラスについて 30 枚とする．観測確率と目標タスクの観測確率の分布がどれだけ近いのかを定量的に評価するため，本稿では χ^2 値を利用する．正規化定数 $N_{vm\lambda}$ あたりのサンプル割合 $\boldsymbol{\eta}$ を $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$ と変化させて，サンプル数による χ^2 値の変化を検証する．また，その際のクラス平均正解率もあわせて評価する．ただし η_0 と η_1 は同じ値とする．各割合で 5 回実行し，その平均で評価する．

表 4 サンプリングの割合を変化させた際の χ^2 値（上）とクラス平均正解率（下）

Table 4 Chi-square score of different sampling rate (above) and mean class accuracy (below).

観測確率	サンプリングの割合					
	0	0.2	0.4	0.6	0.8	1.0
θ_{m0}	2.26	1.27	1.60	2.06	2.43	2.82
θ_{m1}	15.20	5.62	4.29	3.96	3.84	3.91

サンプリングの割合						目標タスクの観測確率
0	0.2	0.4	0.6	0.8	1.0	
.457	.457	.460	.465	.465	.461	.511

表 4 が実験結果である．サンプル数を増やすことで χ^2 値が低くなっていることから，タスク間の観測確率の距離が縮まっていることが分かる．また，それに応じてクラス平均正解率が向上し，目標タスクの観測確率に近くなっていることが確認できる．しかし，割合が 1.0 のように増えすぎると χ^2 値が大きくなり，クラス平均正解率も低下してしまうことが分かった．これは 3.3.3 項で述べたように，事前知識である $p(c_m)$ の影響が大きくなってしまうためと考えられる．

5. 既存研究との比較実験

本章では提案手法を既存研究と比較実験し，提案手法の有効性について議論する．

5.1 節では，AwA を用いた既存研究との比較実験をし，5.2 節で aP-aY を使って既存研究と比較する．そして 5.3 節で，観測確率の近さに関する考察をする．

5.1 実験 1：Animals with Attributes

5.1.1 実験 1(a)：DAP モデルとの比較

ベースライン手法を DAP モデル [4], [10] とし，本稿の提案手法との比較を行った．この実験では提案モデルと DAP モデルでデータセットを AwA とした際のクラス平均正解率で評価した．元タスクと目標タスクのクラスは，それぞれデフォルトで分けられている訓練事例集合とテスト事例集合のクラスラベルの定義域とする．元タスクの訓練事例数は各クラスあたり 10 枚から 90 枚へと 10 枚ずつ増やし，それぞれで学習段階を行った．また，元タスクでの検証用事例数は，元タスクの訓練事例集合の 10%とした．目標タスクのテスト事例集合は各クラスあたり 90 枚で固定し，10 クラス分類を推定する．画像特徴量は DeCAF を使

表 5 ゼロショット学習の既存研究との比較
Table 5 Comparison with the state-of-the-art zero-shot learning.

手法	補助情報	クラス平均正解率（%）
提案手法	属性	46.5（51.1 using the observation probabilities on the target task）
DAP	属性	40.5（文献 [4]）/41.4（文献 [10]）/46.2（our implementation）
IAP	属性	27.8（文献 [4]）/42.2（文献 [10]）
ALE/HLE/AHLE [13]	属性/WordNet/属性と WordNet	37.4/39.0/43.5
Semantic Graph [11]	言語ベクトル（Wikipedia）	43.1
TMV-BLP [12]	属性と言語ベクトル（Wikipedia）	**47.1**

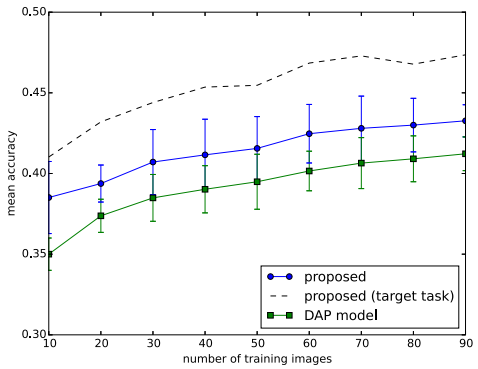

図 6 提案手法と DAP モデルの比較
Fig. 6 Comparison of our proposed method and DAP model.

い，分類器は既存手法と DAP モデルともにパラメータ C を 1.0 とした L2 ロジスティック回帰を用いた．ハイパーパラメータは $\alpha_0 = \alpha_1 = 2$ とした．また，3.3.3 項で示したサンプリングによる方法を採用し，正規化定数 $N_{vm\lambda}$ あたりのサンプル割合 $\boldsymbol{\eta}$ は，$\eta_0 = \eta_1 = 0.2$ とした．元タスクの訓練事例集合は，データセットの訓練事例集合から各クラスあたりランダムに選択するが，選び方の影響を軽減するため，評価指標は各枚数で 5 回実験したそれぞれのクラス平均正解率の平均とする．

クラス平均正解率の比較結果が，**図 6** である．横軸が元タスクの各クラスごとの訓練事例数で縦軸がクラス平均正解率である．青線が提案手法，緑線が DAP モデルを表し，各グラフのエラーバーは 5 回実験した結果の標準偏差を表す．また，提案手法において観測確率を目標タスクの観測確率とした場合の実験結果も点線で表示した．目標タスクの観測確率は，3.3.3 項で述べたように，目標タスクのテスト事例集合にクラスラベルがあるものとして求めた値である．なお目標タスクの観測確率は，実際のゼロショット学習では求めることができないことに留意する．

図 6 より，提案モデルが DAP モデルと比較して正解率が 2〜3%高いことが確認された．さらに，目標タスクの観測確率の場合，それ以上に正解率が高くなることを確認した．目標タスクの観測確率による結果は提案手法の潜在的な精度を示しており，この結果から潜在的にはさらに高い正解率となることが示された．

5.1.2 実験 1(b)：既存のゼロショット学習との比較

本実験では，提案手法と DAP モデル以外の既存のゼロショット学習との比較をする．提案手法と比較する手法は，本稿のベースライン手法である DAP モデルの他，同じく Lampert らによる IAP モデル [4], [10]，Akata らの ALE, HLE，AHLE モデル [13]，Fu らの Semantic Graph [11]，Fu らの TMV-BLP [12] とした．

データセットはすべての手法で AwA とし，元タスク・目標タスクのクラスは，それぞれデフォルトで分けられている訓練事例集合とテスト事例集合のクラスラベルの定義域とし，AwA のすべての訓練事例集合とテスト事例集合を利用する．これらの設定はすべての手法で共通である．本研究の提案モデルでは，元タスクの検証用事例集合は訓練事例集合のうち各クラスにつき 30 枚とし，$C = 1.0$ の L2 ロジスティック回帰を分類器とした．その他のパラメータは $\alpha_0 = \alpha_1 = 2$，$\eta_0 = \eta_1 = 0.8$ とした．提案モデルでは画像特徴量の種類として DeCAF を使っているが，その他の手法では Semantic Graph が同様に DeCAF を用いている以外は HSV color histograms，SIFT [19]，rgSIFT [20]，PHOG [21]，SURF [22]，local self-similarity histograms [23] の 6 種類の画像特徴量を使っている．その他の違いとして，AHLE は属性の他に WordNet の知識を補助情報とし，Semantic Graph は属性の代わりに各クラスの Wikipedia から skip-gram モデルで抽出した言語ベクトルを補助情報としている．さらに TMV-BLP については補助情報として属性と Wikipedia から skip-gram モデルで抽出した言語ベクトルを併用している．このように，そもそも利用した補助情報が異なるうえ，入力とする画像特徴量の種類も異なるので，正解率のみから手法自体の比較を行うことは困難であることに留意されたい．

表 5 がクラス平均正解率で比較した結果である．既存手法の正解率はすべてそれらの論文に書かれている結果を引用し，DAP モデルは本稿で利用した DeCAF での実験結果もあわせて載せた．また，それぞれの手法で使われている補助情報も記載した．表 5 から，これらの手法の中では，TMV-BLP モデルが最も高い正解率であることが確認できる．本稿の提案手法は，TMV-BLP には劣るが，同じ DeCAF 特徴量を用いている Semantic Graph より高い結果となっている．ただし上記のとおり，表 5 の下 3 つの既存手法では，補助情報として属性以外の WordNet や Wikipedia の知識を利用もしくは併用している．このこと

から，提案手法は補助情報について前処理や追加をいっさいしなくても，他のゼロショット学習の手法と同等以上の精度となることを示した．また，提案モデルにおいて，目標タスクの観測確率を用いた場合の正解率は51.1%となり，潜在的には提案モデルは他の手法を上回る精度となることが確認できた．

5.1.3 実験 1(c)：n ショット学習での実験

本実験では，ゼロショットでない問題設定でも提案モデルが有効であるであることを示す．属性ベースゼロショット学習の既存手法である DAP モデルでは，元タスクの分類器を目標タスクで再利用することでゼロショット学習を実現していた．よって，元タスクとは別に分類器を用意して目標タスクの訓練事例集合を学習することはできなかった．これは，DAP モデルがゼロショット学習の場合しか対応していないことを意味する[*6]．一方，提案モデルでは分類器だけではなく，元タスクで学習したパラメータである観測確率も目標タスクで再利用している．よって，たとえそれぞれのタスクで分類器を別々に学習しても，観測確率は再利用できるので，元タスクの知識を目標タスクに移すことができる．すなわち，ゼロショット学習だけではなく，目標タスクで少数の訓練事例集合（たとえば各クラスについて 2 枚）を用意して学習した場合でも適用できる．本稿ではこれを n ショット学習（n は目標タスクの各クラスごとの訓練事例数）と呼ぶ．

本稿では，観測確率によって元タスクの知識を目標タスクに移すことを “転移” と呼ぶ．実験では n ショット学習において，転移しない場合と，提案手法で転移した場合で比較する．データセットは AwA とし，元タスクと目標タスクのクラスは，それぞれデフォルトで分けられている訓練事例集合とテスト事例集合のクラスラベルの定義域とする．転移しない場合は，DAP モデルで目標タスクのみで訓練・推定をし，元タスクの訓練事例数は各クラスについて n 枚，テスト事例数は 80 枚として，10 クラス分類を推定する．転移する場合は提案モデルを用い，目標タスクについては同様の設定とし，元タスクの訓練事例数は各クラスあたり 90 枚，検証用事例数はそのうち 10%とした．転移しない場合と転移した場合の両方で，画像特徴量はこれまでと同様 DeCAF とし，分類器は $C = 1.0$ の L2 ロジスティック回帰とする．転移した場合のパラメータ設定は，$\alpha_0 = \alpha_1 = 2$，$\eta_0 = \eta_1 = 0.2$ とした．評価はクラス平均正解率とし，n を 2 から 10 までの 9 通りの値として，それぞれ 5 回実験した平均とする．$n = 1$ の場合は，目標タスクの訓練事例が各クラス 1 枚しかないことになり，転移しない場合において厳しすぎる設定のため除外した．

実験結果が**図 7** である．横軸が目標タスクの各クラスあたりの訓練事例数，縦軸がクラス平均正解率である．また，

[*6] 逐次学習可能な分類器を利用すれば，ゼロショット以外でも対応可能だが，任意の分類器が使えるという利点は失われる．

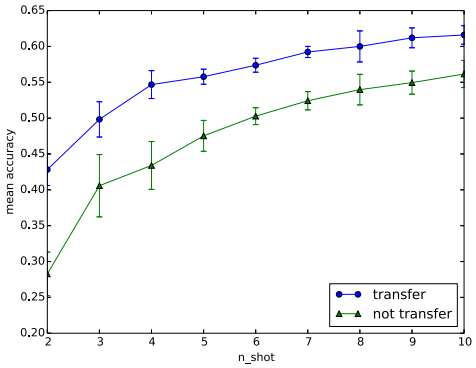

図 7 n ショット学習（転移なしと転移ありの比較）
Fig. 7 n-shot learning (Comparison of “transfer” and “not transfer”).

緑線が転移しない場合，青線が転移した場合を表し，各グラフのエラーバーは標準偏差を表す．提案モデルによって転移した結果の方が高い正解率となっている．特に $n = 2$ のとき，転移しない場合と比べて約 15%正解率が向上していることが確認できた．このように提案手法が DAP モデルができなかった n ショット学習でも有効であることを示した．

5.2 実験 2：aPascal-aYahoo

本実験ではデータセットに aP-aY を用いる．aP-aY は AwA とは異なり，属性が各事例に対して定義されている．よって，本来の属性の定義どおり事例ごとに考えることができるが，AwA のようにクラスごとの定義も考えることができる．これは同じクラスラベルである全事例の属性値を平均して，0 より大きければそのクラスラベルを持つすべての事例の属性値を 1 とすることで求まる．よって，このデータセットでの実験では，事例ごとの定義のほかに，クラスごとの定義の属性値ラベルを考えることができ，前者の実験設定を Per-Image，後者を Per-Class としてそれぞれ実験する．

実験はこれまでの実験と同様，DAP モデルと提案モデルを比較する．画像特徴量の種類はテクスチャやカラー，エッジ，HOG 特徴量による visual word などから作成した 9,751 次元の画像特徴量を利用した．分類器は画像特徴量の種類に合わせて χ^2 カーネルの SVM を利用し，確信度は Platt scaling [24] によって求めた．パラメータ C は 1.0 としパラメータ γ は元タスクの訓練事例集合の χ^2 距離の逆数とした．また，元タスク・目標タスクのクラスラベルの定義域は，それぞれ Pascal の 20 クラス，Yahoo の 12 クラスとした．訓練事例数は各クラスについて $\{10, 20, 30, 40, 50\}$ と変化させ，目標タスクのテスト事例数は各クラスごとに 50 枚とした．以上の設定は比較する両モデルで共通である．提案モデルについては，検証用事例数を訓練事例数の 10%とし，パラメータ設定は，$\alpha_0 = \alpha_1 = 2$，$\eta_0 = \eta_1 = 0.2$ とした．評価はクラス平均正解率で，5 回実行した平均をとる．

5.2.1 実験 2(a)：Per-Class

Per-Class での実験結果が**図 8** である．横軸は元タスクにおける各クラスごとの訓練事例数で縦軸がクラス平均正解率である．青線が提案手法，緑線が DAP モデルを表し，各グラフのエラーバーは標準偏差を表す．また，点線はランダムにクラス分類した際の正解率を表す．提案手法によっておおむね正解率が向上したことが確認できた．しかし，訓練事例数が各クラス 10 枚の際に提案手法が DAP モデルよりも悪くなっている．この理由の 1 つとして，元タスクにおける訓練事例数が少ないことで，モデルのパラメータである観測確率の推定がうまくいかなかった可能性が考えられる．

5.2.2 実験 2(b)：Per-Image

Lampert らによると，Per-Image は画像ごとのクラスを考慮せずに属性を定義していること，そして事例ごとの属性値のラベルは適切に学習できない，という理由によって Per-Class よりもうまく推定できないとされている [10]．Per-Image の実験結果は**図 9** である．グラフについての説明は図 8 と同様である．この結果から，DAP モデルは Per-Class の結果と比較して明らかに適切に分類できておらず，ランダムな分類の正解率とほぼ等しくなっていることが分かる．一方提案手法は，DAP モデルよりも正解率が高く，訓練事例数を増やすことによって徐々に正解率が高くなっていることが分かる．ただし図 9 で，元タスクにおける訓練事例数が 40 枚のとき正解率が下がっていることが確認できる．この現象は DAP モデルの結果でも同様にみられることから，提案モデルの問題ではなく，訓練事例集合もしくは分類器の性質によるものと考えられる．今度は，Per-Class における提案手法の結果と比較すると，同等もしくはそれ以上の正解率となっていることが分かる．このことから，提案手法では DAP モデルの Per-Image では適切に学習できないという問題点を解消し，どちらの問題設定でも同様の性能を得ることができることを示した．この理由としては，提案モデルはうまく学習できなかった分類器を観測確率が考慮し，クラス推定への影響をおさえることができたということが考えられる．

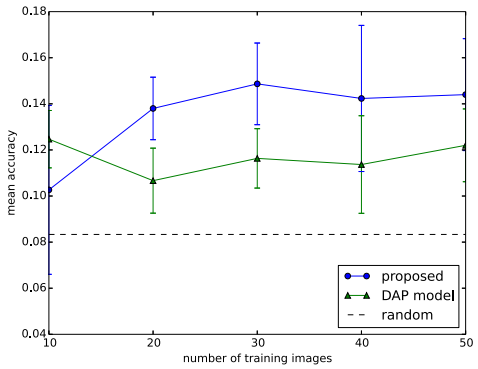

図 8 提案手法と DAP モデルの比較（Per-Class）
Fig. 8 Comparison of our proposed method and DAP model (Per-Class).

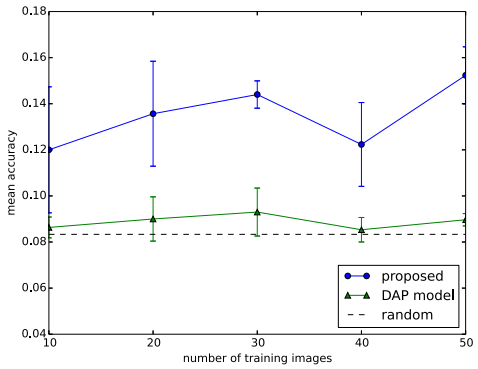

図 9 提案手法と DAP モデルの比較（Per-Image）
Fig. 9 Comparison of our proposed method and DAP model (Per-Image).

5.3 観測確率の近さに関する考察

実験結果によって，提案手法が DAP モデルと比較して正解率が向上することが確認できたが，目標タスクの観測確率で求めた正解率よりも低くなることも明らかになった．本稿では，目標タスクの属性の周辺確率をサンプリングによって事前知識として考慮することで，目標タスクの観測確率の結果に近づけようとした．検証 3 の結果から，この提案手法によって，観測確率が近づき正解率が向上することを確認した．しかし，実験 1(a)，1(b) の結果から，目標タスクの観測確率の結果との差は開いていることが確認できる．ここでは，目標タスクの観測確率にガウス分布によるノイズを与えることで意図的に分布が遠くなるようにし，観測確率の近さと正解率の高さにどこまで相関があるのかを検証する．分布の距離とクラス平均正解率で評価する．

ガウス分布の分散を変更したときのクラス平均正解率の変化が**表 6** であり，χ^2 値の変化が**表 7** である．

表 6 から，分散を大きくすると正解率が低くなっていることが分かる．また表 7 から，分散を大きくすると χ^2 値が大きくなることが確認できる．一方実験 1(a) などではサンプル割合 $\boldsymbol{\eta}$ を $\eta_0 = \eta_1 = 0.2$ としており，このときの χ^2 値は表 4 から確認できる．しかし，この値は表 7 での分散 0.05 のときの χ^2 値よりも小さい値である．さらに表 6 よ

表 6 目標タスクの観測確率にノイズを加えた際のクラス平均正解率
Table 6 Class mean accuracy of different noise to the correct observation probability.

手法	各クラスの訓練事例数				
	10	30	50	70	90
目標タスクの観測確率	.411	.445	.454	.473	.475
$\sigma^2 = 0.05$	.388	.445	.452	.458	.466
$\sigma^2 = 0.1$	.335	.409	.430	.441	.450
$\sigma^2 = 0.15$	.335	.368	.394	.411	.426
DAP モデル	.354	.387	.401	.412	.417

表 7 目標タスクの観測確率にノイズを加えた際の χ^2 値
Table 7 Chi-square score of different noise to the correct observation probability.

観測確率	ノイズ	各クラスの訓練事例数				
		10	30	50	70	90
θ_{m0}	$\sigma^2 = 0.05$	1.61	2.00	2.41	2.57	2.71
	$\sigma^2 = 0.1$	4.00	5.00	5.73	6.06	6.35
	$\sigma^2 = 0.15$	7.52	9.63	10.85	11.47	11.97
θ_{m1}	$\sigma^2 = 0.05$	3.73	3.88	4.33	4.95	4.75
	$\sigma^2 = 0.1$	5.01	5.27	5.70	6.45	6.16
	$\sigma^2 = 0.15$	6.92	7.37	7.75	8.62	8.23

り，分散 0.05 のときの正解率は目標タスクの観測確率のときとほぼ同じであることから，χ^2 値から判断すれば，提案手法でも目標タスクの観測確率の正解率とほぼ同じになるはずである．しかし，実験 1(a) でも示したように，実際には目標タスクの観測確率の方が高い正解率となっている．したがって，目標タスクの観測確率の場合と比べて提案手法の正解率が低いことの原因は，χ^2 値で分かるような観測確率全体を分布とした場合の違いだけではないことが示された．本稿では，大まかに近さを測る手段として χ^2 値を採用したが，他の評価方法で検証する，または観測確率の差が大きくなると結果に影響されるような属性を見つける，といった工夫が必要と考えられる．

6. 結論

6.1 結果

本稿では，属性ベース転移学習において，これまで着目されていなかった各属性の画像特徴量への現れやすさに着目した．そしてこの度合いを観測確率として考慮した新しい属性ベースゼロショット学習のモデルを提案した．モデルを提案するにあたり，いくつかの仮定をしたが，それぞれについて検証を行い，おおむね妥当であることを示した．また異なるタスクでの観測確率を近づけるために，目標タスクでの属性の周辺確率をサンプリングによって観測確率の事前知識とすることで，タスク間の観測確率をより近づけることが可能であることを示した．

実験では，提案モデルが DAP モデルよりも良い正解率となることを示した．また，既存のゼロショット学習とも比較し，これらの研究が別の知識を補助情報として精度を上げようとするなか，本稿の提案手法はデータに関する前処理をいっさいせずに同等以上の結果になることを確認した．さらに DAP モデルではできなかった n ショット学習を行い，元タスクからの転移によって，正解率が向上することを確認した．また，属性が事例ごとに定義されている場合に DAP モデルがうまく推定できないことが知られていたが，提案モデルによってこの問題が解消されることが確認できた．このような結果から，本稿の提案手法は属性ベースゼロショット学習のモデルとして有効であることが示された．

一方で，観測確率を目標タスクの観測確率に近づけても，本研究の結果では目標タスクの観測確率による理想的な正解率に到達できなかったことが確認された．また観測確率に関する考察から，この原因は，観測確率全体の分布の違いだけではないことが示された．

6.2 今後の展望

本稿では，観測確率によって良い分類結果を得られることを複数の実験結果から示したが，今後の課題として，提案手法がどのような場合に有効なのかを実験および理論を通して検証する必要がある．特に観測確率がどのような場合に良い推定を得られるかは興味深い課題であり，本稿で検証した χ^2 値以外の評価方法で検証するほか，様々なデータセットでの検証実験に取り組みたい．

また，本稿では属性ベースのゼロショット学習に取り組んだが，この分野の共通の課題として，属性をどのように定義するかということがあげられる．本稿の提案手法の強みは，属性の観測確率を考慮することで，クラス推定に役立たない属性の重要度を下げることができるということである．よって，属性の定義を人手で行った場合，既存手法と比較して提案手法の大幅な分類精度の改善が期待できる．ここで重要なのが，本稿では属性の重要度，すなわち観測確率を学習段階で求めることができるという点である．よって，前処理をせずに得られた属性をそのまま使った場合でも，クラスの推定が可能と考えられる．

さらに，本稿で提案した手法は生成モデルでありながら特定の分類器によらないアルゴリズムである．よって今後は画像以外のメディア情報でも提案モデルが有効かどうか検証したい．

参考文献

[1] Biederman, I.: Recognition-by-components: A theory of human image understanding, *Psychological Review*, Vol.94, pp.115–147 (1987).
[2] Pan, S.J. and Yang, Q.: A survey on transfer learning, *IEEE Trans. Knowledge and Data Engineering*, Vol.22, No.10, pp.1345–1359 (2010).
[3] Larochelle, H., Erhan, D. and Bengio, Y.: Zero-data Learning of New Tasks, *Proc. National Conference on Artificial Intelligence*, pp.646–651 (2008).
[4] Lampert, C.H., Nickisch, H. and Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.951–958 (2009).
[5] Yu, X. and Aloimonos, Y.: Attribute-based transfer learning for object categorization with zero/one training example, *Proc. European Conference on Computer Vision*, Vol.6315, pp.127–140 (2010).
[6] Rohrbach, M. and Stark, M.: What helps where - and why? Semantic relatedness for knowledge transfer, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.910–917 (2010).

[7] Rohrbach, M., Stark, M. and Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1641–1648 (2011).
[8] Socher, R. and Ganjoo, M.: Zero-shot learning through cross-modal transfer, *CoRR*, Vol.abs/1301.3666, pp.1–7 (2013).
[9] Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S. and Dean, J.: Zero-Shot Learning by Convex Combination of Semantic Embeddings, *CoRR*, Vol.abs/1312.5650, pp.1–9 (2013).
[10] Lampert, C.H., Nickisch, H. and Harmeling, S.: Attribute-based classification for zero-shot visual object categorization, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.36, No.3, pp.453–465 (2014).
[11] Fu, Z.-Y., Xiang, T. and Gong, S.: Semantic Graph for Zero-Shot Learning, *CoRR*, Vol.abs/1406.4112, pp.1–9 (2014).
[12] Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z. and Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation, *Proc. European Conference on Computer Vision*, Vol.8690, pp.584–599 (2014).
[13] Akata, Z., Perronnin, F., Harchaoui, Z. and Schmid, C.: Label-embedding for attribute-based classification, *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.819–826 (2013).
[14] Parikh, D. and Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1681–1688 (2011).
[15] Suzuki, M., Sato, H., Oyama, S. and Kurihara, M.: Image Classification by Transfer Learning Based on the Predictive Ability of Each Attribute, *Proc. International MultiConference of Engineers and Computer Scientists*, pp.75–78 (2014).
[16] Suzuki, M., Sato, H., Oyama, S. and Kurihara, M.: Transfer learning based on the observation probability of each attribute, *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, pp.3648–3652 (2014).
[17] Farhadi, A., Endres, I., Hoiem, D. and Forsyth, D.: Describing objects by their attributes, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1778–1785 (2009).
[18] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition, *International Conference on Machine Learning*, Vol.32, pp.647–655 (2014).
[19] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
[20] van de Sande, K., Gevers, T. and Snoek, C.: Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.32, No.9, pp.1582–1596 (2010).
[21] Bosch, A., Zisserman, A. and Munoz, X.: Representing Shape with a Spatial Pyramid Kernel, *Proc. ACM International Conference on Image and Video Retrieval*, pp.401–408 (2007).
[22] Bay, H., Ess, A., Tuytelaars, T. and Van Gool, L.: Speeded-Up Robust Features (SURF), *Computer Vision and Image Understanding*, Vol.110, No.3, pp.346–359 (2008).
[23] Shechtman, E. and Irani, M.: Matching Local Self-Similarities across Images and Videos, *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2007).
[24] Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, *Advances in Large Margin Classifiers*, pp.61–74 (1999).

付　　録

A.1 観測確率の導出

事後確率 $p(\theta_{m\lambda}|\mathbf{X})$ が最大となるような観測確率の推定量 $\hat{\theta}_{m\lambda}$ を求める．

$p(\theta_{m\lambda}|\mathbf{X})$ について対数をとると，

$$
\begin{aligned}
&\log p(\theta_{m\lambda}|\mathbf{X}) \\
&\propto \log p(\mathbf{X}|\theta_{m\lambda}) + \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&= \sum_{i:a_{mi}=\lambda} \log p(\mathbf{x}_i|\theta_{m\lambda}) + \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&= \sum_{i:a_{mi}=\lambda} \log \sum_{c_{mi}} p(\mathbf{x}_i, c_{mi}|\theta_{m\lambda}) + \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&= \sum_{i:a_{mi}=\lambda} \log \sum_{a} p(c_{mi}|\mathbf{x}_i) \frac{p(\mathbf{x}_i, c_{mi}|\theta_{m\lambda})}{p(c_{mi}|\mathbf{x}_i)} + \log p(\theta_{m\lambda}|\boldsymbol{\alpha})
\end{aligned}
$$

となる．ここで $\log(x)$ は凸関数なので，Jensen の不等式を適用すると

$$
\begin{aligned}
&\sum_{i:a_{mi}=\lambda} \log \sum_{a} p(c_{mi}|\mathbf{x}_i) \frac{p(\mathbf{x}_i, c_{mi}|\theta_{m\lambda})}{p(c_{mi}|\mathbf{x}_i)} + \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&\geq \sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i) \log \frac{p(\mathbf{x}_i, c_{mi}|\theta_{m\lambda})}{p(c_{mi}|\mathbf{x}_i)} \\
&\quad + \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&= \sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i) \{\log p(\mathbf{x}_n|c_{mi}) \\
&\qquad\qquad + \log p(c_{mi}|\theta_{m\lambda})\} \\
&\quad - \sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i) \log p(c_{mi}|\mathbf{x}_i) \\
&\quad + \log p(\theta_{m\lambda}|\boldsymbol{\alpha})
\end{aligned}
$$

となり，下界が求まる．

下界を $\theta_{m\lambda}$ について最大化するために偏微分をとる．

$$
\begin{aligned}
&\sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i) \frac{\partial}{\partial \theta_{m\lambda}} \log p(c_{mi}|\theta_{m\lambda}) \\
&\quad + \frac{\partial}{\partial \theta_{m\lambda}} \log p(\theta_{m\lambda}|\boldsymbol{\alpha}) \\
&= \sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i) \\
&\qquad \times \frac{\partial}{\partial \theta_{m\lambda}} \log \{{\theta_{m\lambda}}^{c_{mi}} (1-\theta_{m\lambda})^{1-c_{mi}}\} \\
&\quad + \frac{\partial}{\partial \theta_{m\lambda}} \log \left\{ \frac{1}{Beta(\alpha_0, \alpha_1)} {\theta_{m\lambda}}^{\alpha_0 - 1} (1-\theta)^{\alpha_1 - 1} \right\} \\
&= \sum_{i:a_{mi}=\lambda} \sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i)
\end{aligned}
$$

$$
\begin{aligned}
&\times \frac{\partial}{\partial \theta_{m\lambda}}\{c_{mi}\log\theta_{m\lambda}+(1-c_{mi})\log(1-\theta_{m\lambda})\}\\
&+\frac{\partial}{\partial \theta_{m\lambda}}\{-\log Beta(\alpha_0,\alpha_1)+(\alpha_0-1)\log\theta_{m\lambda}\\
&\qquad +(\alpha_1-1)\log(1-\theta_{m\lambda})\}\\
=&\sum_{i:a_{mi}=\lambda}\sum_{c_{mi}} p(c_{mi}|\mathbf{x}_i)\left\{\frac{c_{mi}}{\theta_{m\lambda}}-\frac{1-c_{mi}}{1-\theta_{m\lambda}}\right\}\\
&+\frac{\alpha_0-1}{\theta_{m\lambda}}-\frac{\alpha_1-1}{1-\theta_{m\lambda}}\\
=&\sum_{i:a_{mi}=\lambda}\left\{\frac{p(c_{mi}=1|\mathbf{x}_i)}{\theta_{m\lambda}}-\frac{p(c_{mi}=0|\mathbf{x}_i)}{1-\theta_{m\lambda}}\right\}\\
&+\frac{\alpha_0-1}{\theta_{m\lambda}}-\frac{\alpha_1-1}{1-\theta_{m\lambda}}\\
=&\sum_{i:a_{mi}=\lambda}\left\{\frac{p(c_{mi}=1|\mathbf{x}_i)}{\theta_{m\lambda}}-\frac{1-p(c_{mi}=1|\mathbf{x}_i)}{1-\theta_{m\lambda}}\right\}\\
&+\frac{\alpha_0-1}{\theta_{m\lambda}}-\frac{\alpha_1-1}{1-\theta_{m\lambda}}
\end{aligned}
$$

この式が 0 となるような $\theta_{m\lambda}$ を $\hat{\theta}_{m\lambda}$ とすると，$\hat{\theta}_{m\lambda}$ は

$$\hat{\theta}_{m\lambda}=\frac{\sum_{i:a_{mi}=\lambda}p(c_{mi}=1|\mathbf{x}_i)+\alpha_0-1}{\sum_{i:a_{mi}=\lambda}+\alpha_0+\alpha_1-2}$$

となる．

鈴木 雅大

2013 年北海道大学工学部情報エレクトロニクス学科卒業．2015 年同大学大学院修士課程修了．同年東京大学大学院工学系研究科博士課程入学．人工知能，機械学習の研究に従事．

佐藤 晴彦 （正会員）

2005 年北海道大学工学部情報工学科卒業．2007 年同大学大学院修士課程修了．2008 年同博士課程修了．情報科学博士．2009 年北海道大学助教．定理自動証明の研究に従事．電子情報通信学会，日本ソフトウェア科学会各会員．

小山 聡 （正会員）

1994 年京都大学工学部数理工学科卒業，1996 年同大学大学院工学研究科修士課程修了．日本電信電話株式会社，京都大学大学院情報学研究科博士後期課程，日本学術振興会特別研究員（DC），京都大学大学院情報学研究科助手，スタンフォード大学 Visiting Assistant Professor 等を経て，2009 年より北海道大学大学院情報科学研究科准教授．博士（情報学）．主な研究分野は機械学習，データマイニング，情報検索，クラウドソーシング等．2005 年度人工知能学会論文賞，2009 年度日本データベース学会上林奨励賞受賞．

栗原 正仁 （正会員）

1955 年生．1978 年北海道大学工学部卒業．1980 年同大学大学院工学研究科情報工学専攻修士課程修了．工学博士．現在，北海道大学大学院情報科学研究科情報理工学専攻教授．ソフトウェア科学および人工知能の境界領域の研究に興味を持つ．1990 年情報処理学会創立 25 周年記念論文賞，2011 年電子情報通信学会論文賞受賞．電子情報通信学会，日本ソフトウェア科学会，人工知能学会，日本知能情報ファジィ学会各会員．

松尾 豊 （正会員）

1997 年東京大学工学部卒業．2002 年同大学大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より東京大学大学院工学系研究科技術経営戦略学専攻准教授．2012 年より人工知能学会理事・編集委員長，2014 年より倫理委員長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞等受賞．専門は，Web 工学，Deep Learning，人工知能．