Vol.2015-NL-222 No.4
2015/7/15

Zero-shot learning における線形回帰の影響

重藤 優太郎[1,a)]　鈴木 郁美[3,b)]　原 一夫[,c)]　新保 仁[1,d)]　松本 裕治[1,e)]

概要：本稿では，線形回帰を用いた zero-shot learning におけるハブの影響について議論する．特に，事例空間とラベル空間の関係に注目する．これまでの先行研究は，事例空間からラベル空間への写像を提案していたが，本稿では，これとは逆にラベル空間から事例空間への写像を考える．素朴なデータモデルを考えることによって，この逆向きの写像が，ZSL のラベル予測（近傍検索）においてハブの出現を抑制することを示す．また，実データを用いて，その検証を行う．

1. はじめに

1.1 背景

Drug discovery [12] や対訳抽出 [7], [8], [17], 画像ラベル付け（image labeling） [2], [9], [18], [19], [22] などの多くのタスクは zero-shot learning (ZSL) [12] のタスクとして定式化できる．ZSL は訓練セットとして事例とラベルのペアの集合が与えられ，それを用いてラベルが未知の評価事例に対して分類を行う．この点で，ZSL は分類問題の一種とみなせる．ただし，一般的な分類問題は，評価事例のラベルが，与えられた訓練セットに含まれていることを仮定しているが，ZSL ではこの仮定を置いておらず，加えて，ラベル数が膨大である場合が多い．例えば，原言語の単語の翻訳単語を予測する問題である対訳抽出では，ラベル数は原言語の単語の翻訳単語の数である．

当然ながら，評価事例のラベルが訓練セットに含まれていない場合，一般的な分類問題と同様の解法を取ることはできない．このため，ZSL はラベルが距離空間（ラベル空間）に埋め込まれており，この空間においてラベル間距離（もしくは類似度）が計算できると仮定している．このラベル空間は，背景知識や外部知識などを用いて構築できる．例えば，画像ラベル付け問題では，ラベルは画像に付与するキーワードであり，このキーワードのベクトル表現は大量のテキストから学習される．

ラベル空間が構築された後は，回帰と近傍検索を用いることによってラベルの予測を行う．すなわち，事例が点として表現されている空間（事例空間）からラベル空間への写像関数を学習し，学習した写像関数によって評価事例をラベル空間に写像する．写像された評価事例と距離が最も近いラベルを検索し，そのラベルを最終的な予測結果とする．

事例空間からラベル空間への写像はリッジ回帰を用いる手法 [7], [8], [17], [19] やニューラルネットワークを利用する手法 [9], [18], [22] が提案されている．

一方，近年になって，高次元空間における近傍検索の問題として「ハブの出現」 (hubness phenomenon) [20] が注目されている．ハブとは「多数の事例の近傍事例になる事例」を指す．近傍検索は与えられたクエリ事例と，距離が小さい事例をデータセット中から検索するタスクである．この際, データ中にハブが存在すると，どんなクエリを与えても，ほとんどの場合ハブ事例が検索結果に含まれる．そのような検索結果は，有意義な結果であるとは言えず，実用上役に立たない．

1.2 研究目的・貢献

本稿では，ZSL における回帰と近傍検索の関係について議論を行う．

Dinu and Baroni [8] は，リッジ回帰によって事例をラベル空間に写像した場合，近傍検索において，少数のラベルがハブとなり，結果として，ZSL の予測精度が悪化することを報告した．

本研究の目的はハブの出現を抑制し，ZSL の予測精度を改善することである．本研究の貢献を以下に示す．

- 本稿では，ZSL においてリッジ回帰（もしくは最小二乗法）を写像関数として用いた場合におけるハブの出現するメカニズムを示した．その分析の結果として，

[1] 奈良先端科学技術大学院大学 奈良県生駒市高山町 8916-5
[2] 統計数理研究所 統計的機械学習センター
[3] 国立遺伝学研究所 生命情報研究センター
[a)] yutaro-s@is.naist.jp
[b)] suzuki.ikumi@gmail.com
[c)] kazuo.hara@gmail.com
[d)] shimbo@is.naist.jp
[e)] matsu@is.naist.jp

データの次元数だけではなく，上で述べたリッジ回帰の使い方ではハブの出現を促進してしまうことがわかった．

- 上記の分析より，本稿では，リッジ回帰によってラベルを事例空間に写像することを提案する．先行研究では，事例をラベル空間に写像していたが，本提案手法は，これとは逆方向の写像になっている．
  6 節の実験結果で示すが，人工データ・実データの双方で，提案手法は先行研究の予測精度を上回った．
- ハブ研究に対する貢献として，クエリ事例とデータ事例のそれぞれが異なる分布に従っていると仮定した場合，ハブの出現に，データ分布の分散が寄与することを示した．特に，データに存在するバイアスを分散の関数として表現した．

2. リッジ回帰を用いた ZSL

はじめに，X で事例集合，Y でラベル集合を表現する．事例とラベルはベクトルで表現されていると仮定する．

ここで，$X \subset \mathbb{R}^c$ と $Y \subset \mathbb{R}^d$ を定義し，これらの空間 $\mathbb{R}^c$，$\mathbb{R}^d$ をそれぞれ事例空間，ラベル空間と呼ぶ．

訓練事例集合 $X_{\text{train}} = \{\mathbf{x}_i \mid i = 1, \ldots, n\}$ と対応するラベル集合（訓練ラベル集合）$Y_{\text{train}} = \{\mathbf{y}_i \mid i = 1, \ldots, n\}$ が与えられた場合を考える．一般的な分類問題の場合，訓練ラベルの種類はラベル集合と同じであるが（$Y_{\text{train}} = Y$），ZSL では Y_{train} は Y の部分集合となり，評価事例のラベルは Y_{train} に含まれない．すなわち，$Y \backslash Y_{\text{train}}$ である．

このような問題設定の下，評価事例 $\mathbf{x} \in X$ から Y に存在するラベルを直接予測する関数 f を学習することは困難である．そのため，ZSL の多くの研究 [2], [9], [18], [19], [22] が間接的にラベルを予測する．具体的には，回帰によって写像関数 $m : \mathbb{R}^c \to \mathbb{R}^d$ を学習し，この写像関数 m によって写像された $m(\mathbf{x})$ に最も距離が近いラベルを予測結果とする．予測関数 f は以下のように書き表される．

$$f(\mathbf{x}) = \arg\min_{\mathbf{y} \in Y} \|m(\mathbf{x}) - \mathbf{y}\|.$$

評価事例 $\mathbf{x}$ を m によって写像した後は, このタスクはラベル空間における，近傍検索問題となる．

3. ハブの出現とデータの分散

近傍検索を行う場合，常に同一事例が検索結果となることは，望ましくない．Radovanović ら [20] は高次元空間において，このようなハブと呼ばれる事例が出現する理論的背景を示している．ハブの存在は直感的には想像しにくいが，実際に多くの高次元 データセット（人工データ・実データの双方）において，ハブの出現が報告されている [20], [21], [23].

本研究の目的は ZSL の近傍検索におけるハブの出現を分析することである．一方で，既存のハブに関する理論 [20] は異なる次元におけるハブの出現しやすさについてのみ議論しているため，この理論を ZSL の分析に直接用いることはできない．

本研究では，同じ次元だが分散の異なる 2 種類の分布を考え，どちらの分布がよりハブが出現しやすいかを議論する．この議論を行うために，まず以下に補題 1 を示す．この補題は近傍検索に関する補題であり，Radovanović らの定理 [20] と似ているが，クエリ事例とデータ事例が異なる分布に従うと考えている点が異なる．

クエリ事例を $\mathbf{x}$，検索対象となるデータ事例を $\mathbf{y}_1$，$\mathbf{y}_2$ で定義した場合，2 節で述べた ZSL の定式化に当てはめると，$\mathbf{x}$ はラベル空間に（写像関数 m によって）写像された評価事例であり，$\mathbf{y}_1$，$\mathbf{y}_2$ は原点から異なる距離にある評価ラベルとなる．ここで，$\mathbf{x}$ が平均が零の分布 $\mathcal{X}$ からサンプリングされた場合，我々が知りたいことは，$\mathbf{y}_1$ と $\mathbf{y}_2$ のどちらがより $\mathbf{x}$ と距離が近くなりやすいか，また，どの程度なりやすいかということである．

$\mathrm{E}[\cdot]$ と $\mathrm{Var}[\cdot]$ はそれぞれ期待値と分散を表し，また，$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均 $\boldsymbol{\mu}$，共分散 $\boldsymbol{\Sigma}$ の多変量正規分布を表す．

補題 1 $\mathbf{y} = [y_1, \ldots, y_d]^{\mathrm{T}}$ を d 次元のベクトルとし，ベクトルの各要素が互いに独立で $\mathcal{N}(0, s^2)$ に従うとする．すなわち，$\mathcal{Y} = \mathcal{N}(\mathbf{0}, s^2\mathbf{I})$ としたとき，$\mathbf{y} \sim \mathcal{Y}$ である．さらに，$\sigma = \sqrt{\mathrm{Var}_{\mathcal{Y}}[\|\mathbf{y}\|^2]}$ を二乗ノルム $\|\mathbf{y}\|^2$ の標準偏差とする．
ここで，二乗ノルムが $\gamma\sigma$ だけ離れている，2 個の固定点 $\mathbf{y}_1$，$\mathbf{y}_2$ を考える．すなわち，

$$\|\mathbf{y}_2\|^2 - \|\mathbf{y}_1\|^2 = \gamma\sigma$$

となる．$\mathbf{x}$ が零平均の分布 $\mathcal{X}$ からサンプルされたとする．このとき，$\mathbf{x}$ から $\mathbf{y}_1$ への二乗距離と $\mathbf{x}$ から $\mathbf{y}_2$ への二乗距離の差の期待値 Δ，すなわち

$$\Delta = \mathrm{E}_{\mathcal{X}}\left[\|\mathbf{x} - \mathbf{y}_2\|^2\right] - \mathrm{E}_{\mathcal{X}}\left[\|\mathbf{x} - \mathbf{y}_1\|^2\right] \tag{1}$$

は

$$\Delta = \sqrt{2}\gamma d^{1/2} s^2 \tag{2}$$

で与えられる．

補題 1 より，Δ は $\mathbf{x}$ から $\mathbf{y_1}$，$\mathbf{y_2}$ の二乗距離の差の期待値を示している．式 (2) より，γ が増加するとともに，Δ の値も増加する．これは，$\|\mathbf{y}_1\|^2 < \|\mathbf{y}_2\|^2$ である際に，$\mathbf{y}_1$ は $\mathbf{y}_2$ よりも，$\mathcal{X}$ からサンプリングされたクエリと距離が近くなりやすいことを示している．これは，データセット中の任意のデータ事例間で成り立つ．結果として，原点に近い（ノルムが小さい）データ事例がハブになりやすい．

さて，固定 d 次元空間上の固定された分布 $\mathcal{X}$ の下で，分散が異なる二つの分布 $\mathcal{Y}_1$，$\mathcal{Y}_2$ を考え，どちらがよりハブが出現しやすいかという議論をしたい．しかしながら，Δ

の大小のみをもって，$\mathcal{Y}_1$ の方がよりハブが出現しやすいと結論づけることはできない．なぜなら，$\mathcal{Y}$ の分布が異なれば，二乗距離 $\|\mathbf{x}-\mathbf{y}\|^2$ の分布も異なるからである．

そこで，Δ が $\|\mathbf{x}-\mathbf{y}\|^2$ の期待値と比較して，どの程度の影響を持つかについて，以下が示せる．

定理 1 クエリの分布 $\mathcal{X}$ と，データ事例 $\mathbf{y}$ に対して 2 種類の分布 $\mathcal{Y}_1=\mathcal{N}(\mathbf{0},s_1^2\mathbf{I})$ と $\mathcal{Y}_2=\mathcal{N}(\mathbf{0},s_2^2\mathbf{I})$ を考える．ここで，$s_1^2<s_2^2$，また，$\mathcal{X}$ と $\mathcal{Y}$ が独立だと仮定した場合，以下の関係を得る．

$$\frac{\Delta(\gamma,d,s_1)}{\mathrm{E}_{\mathcal{X}\mathcal{Y}_1}[\|\mathbf{x}-\mathbf{y}\|^2]}<\frac{\Delta(\gamma,d,s_2)}{\mathrm{E}_{\mathcal{X}\mathcal{Y}_2}[\|\mathbf{x}-\mathbf{y}\|^2]}. \tag{3}$$

ここでは，Δ を γ，d，s の関数として明記している．

式 (3) をハブの出現しやすさだと考えると，ハブがより出現しにくいという意味で，$\mathcal{Y}$ の分散 s^2 は，より小さい方が望ましい．

4. 回帰を用いた ZSL におけるハブの出現

この節では，ZSL の近傍探索ステップにおけるハブの出現について議論する．この議論によって，リッジ回帰を用いて事例をラベル空間に写像する，既存の ZSL の定式化が，結果的に，ハブの出現を促進してしまうことを示す．以下では，この手法を単に既存法と呼ぶ．

さらにその解決策として，リッジ回帰の写像を逆方向にすることを提案する．すなわち，提案法ではラベルを事例空間に写像し，事例空間において近傍検索を行う．この提案法の正当性を示すために，以下の 3 種類の分析を行う．

(1) リッジ回帰（もしくは線形回帰）は，説明変数を対応する目的変数よりも原点に近い位置に写像する傾向にある．既存法は，事例をラベル空間へ写像するので，写像された事例のノルムはラベルのノルムよりも小さくなる（4.1 節）．

(2) 3 節の結果と上述した結果より，ハブの出現という観点から述べると，（近傍検索の際にクエリとなる）事例がラベルよりも原点の近く位置することは好ましくなく，提案法のように，事例よりもラベルが原点の近くに位置することが望ましい（4.2 節）．

(3) (2) とは独立した議論として，事例とラベルの近傍関係に注目した分析を行う（4.3 節）．この分析の結果，事例よりもラベルが原点の近くに位置することで，正解のラベルが任意の事例の最近傍点となる確率が高くなることがわかる．この分析は，ハブ現象と直接関連がないものの，提案法の正当性を補完する．

4.1 回帰によるノルムの変化

ここでは，リッジ回帰を写像関数として用いることによって，説明変数が目的変数よりも原点に近い位置に写像される傾向にあることを示す．正則化を考えた場合，正則化項によって推定される係数が小さくなるので，この傾向はある程度明らかである．しかしながら，この傾向は正則化項のない最小二乗法においても生じることを示す．

$\|\cdot\|_{\mathrm{F}}$ と $\|\cdot\|_2$ をそれぞれ行列のフロベニウスノルムと 2-ノルムとする．

定理 2 リッジ回帰の写像行列を $\mathbf{M}\in\mathbb{R}^{d\times c}$，説明変数の行列 $\mathbf{A}\in\mathbb{R}^{c\times n}$，目的変数の行列 $\mathbf{B}\in\mathbb{R}^{d\times n}$ と定義した場合，写像行列は

$$\mathbf{M}=\arg\min_{\mathbf{M}}\left(\|\mathbf{M}\mathbf{A}-\mathbf{B}\|_{\mathrm{F}}^2+\lambda\|\mathbf{M}\|_{\mathrm{F}}\right). \tag{4}$$

によって求まる．ここで，$\lambda\geq 0$ は正則化パラメータである．このとき，写像された説明変数と目的変数の関係は $\|\mathbf{M}\mathbf{A}\|_2\leq\|\mathbf{B}\|_2$ である．

本稿では，データが中心化されていることを想定しているので，行列の 2-ノルムは主成分方向の分散を示す尺度として解釈できる．従って，定理 2 は，写像された説明変数 $\mathbf{M}\mathbf{A}$ の主成分方向の分散が，目的変数 $\mathbf{B}$ の分散よりも小さくなりやすいことを示している．

さらに，正則化項が無い場合（$\lambda=0$）においても，$\|\mathbf{M}\mathbf{A}\|_2\leq\|\mathbf{B}\|_2$ は成り立ち，射影された説明変数の分散が小さくなりやすいという傾向が生じる．その結果，単純に正則化パラメータ $\lambda=0$ としても，写像された説明変数の分散が小さくなるという傾向を完全に排除することはできない．

既存法は，$\mathbf{A}$ を訓練事例集合の行列 $\mathbf{X}=[\mathbf{x}_1\cdots\mathbf{x}_n]\in\mathbb{R}^{c\times n}$，$\mathbf{B}$ を訓練ラベル行列 $\mathbf{B}=\mathbf{Y}=[\mathbf{y}_1\cdots\mathbf{y}_n]\in\mathbb{R}^{d\times n}$ として，$\mathbf{A}$ を（写像行列 $\mathbf{M}$ によって）ラベル空間へ写像する．上述した定理 2 より，$\mathbf{A}$ は $\mathbf{B}$ よりも原点に近い位置に写像される．定理 2 は訓練セットのみについて議論しているものの，写像された（$\mathbf{X}$ に含まれない）評価事例も多くの（$\mathbf{Y}$ に含まれる）訓練ラベルよりも，原点に近い位置に写像される傾向にあることを示唆している．

4.2 近傍検索におけるノルムの変化の影響

4.1 節で述べた通り，リッジ回帰は説明変数を目的変数よりも原点に近い位置に写像する傾向がある．既存法は事例 X をラベル空間に写像するので，写像された事例のノルムはラベルのノルムよりも小さくなる傾向にある．

4 節で述べた提案法は，既存法の写像とは逆方向であり，ラベル Y を事例空間へ写像する．すなわち，写像されたラベルのノルムは事例のノルムよりも小さくなることが予想される．

ここで，ZSL の近傍探索ステップにおいて，上記した 2 方向の写像（既存法と提案法）のどちらがより適しているかを考える．本稿では次の仮定の下で，その答えを与える．

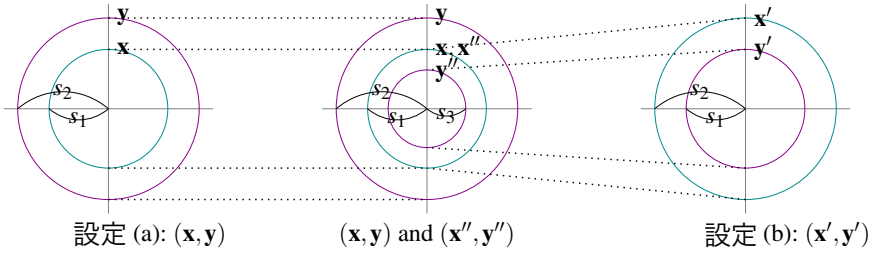

図 1 4.2 節の概略図．左の図と右の図はそれぞれ設定 (a) と (b) を表現している．中央の図は設定 (a) とスケールした設定 (b) を比べた場合を表している．円は分布を表現しており，その半径は標準偏差を示している．$\mathbf{x}$ と $\mathbf{y}'$ の円の半径は s_1 であり，$\mathbf{y}$ と $\mathbf{x}'$ の円の半径は s_2 となる．$\mathbf{x}''$ と $\mathbf{y}''$ はスケールした $\mathbf{x}'$ と $\mathbf{y}'$ であり，$\mathbf{x}''$ の標準偏差は $\mathbf{x}$ と等しく，$\mathbf{y}''$ の標準偏差は $s_3 = s_1^2/s_2$ である．

(i) 事例空間とラベル空間の次元数が同じである．
(ii) 事例とラベルは等方性（isotropically）正規分布に従っている．
(iii) 射影されたデータも同様に等方性（isotropically）正規分布に従っている．

$\mathcal{D}_1 = \mathcal{N}(\mathbf{0}, s_1^2\mathbf{I})$ と $\mathcal{D}_2 = \mathcal{N}(\mathbf{0}, s_2^2\mathbf{I})$ を 2 種類の多変量正規分布と定義する．ここで，$s_1^2 < s_2^2$ とする．このとき，事例 $\mathbf{x}$（既存法においては，その写像）とラベル $\mathbf{y}$（提案法においては，その写像）に関する 2 種類の設定を考える．

(a) $\mathbf{x} \sim \mathcal{D}_1$，$\mathbf{y} \sim \mathcal{D}_2$．
(b) $\mathbf{x}' \sim \mathcal{D}_2$，$\mathbf{y}' \sim \mathcal{D}_1$．

設定 (b) のダッシュは二種類の設定を区別するために用いている．

設定 (a) は既存法，設定 (b) は提案法をモデル化している．より詳細に述べると，設定 (a) は，$\mathbf{x}$ のノルムが $\mathbf{y}$ のノルムよりも小さくなっており，既存法の写像後の状況を模倣している．一方で，設定 (b) は，提案法を模倣しており，$\mathbf{y}$ のノルムが $\mathbf{x}$ のノルムよりも小さくなっている．

今，どちらの設定がよりハブが出現しやすいかを検証する．まず，設定 (b) を (s_1/s_2) によってスケールさせる．すなわち，$\mathbf{x}'' = (s_1/s_2)\mathbf{x}'$，$\mathbf{y}'' = (s_1/s_2)\mathbf{y}'$ とする．ここで，2 変数を同様にスケールしているので，変数間の近傍関係は保存されている．図 1 に $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}'$, $\mathbf{y}'$, $\mathbf{x}''$, $\mathbf{y}''$ の関係を示す．

$\{x_i'\}$ と $\{y_i'\}$ は $\mathbf{x}'$ と $\mathbf{y}'$ の i 番目の要素，同様に，$\{x_i''\}$ と $\{y_i''\}$ は $\mathbf{x}''$，$\mathbf{y}''$ の i 番目の要素とした時，以下の式を得る．

$$\mathrm{Var}[x_i''] = \mathrm{Var}\left[\frac{s_1}{s_2}x_i'\right] = \left(\frac{s_1}{s_2}\right)^2 \mathrm{Var}[x_i'] = s_1^2,$$
$$\mathrm{Var}[y_i''] = \mathrm{Var}\left[\frac{s_1}{s_2}y_i'\right] = \left(\frac{s_1}{s_2}\right)^2 \mathrm{Var}[y_i'] = \frac{s_1^4}{s_2^2}.$$

この式より，$\mathbf{x}''$ は $\mathcal{N}(\mathbf{0}, s_1^2\mathbf{I})$ に従い，$\mathbf{y}''$ は $\mathcal{N}(\mathbf{0}, (s_1^4/s_2^2)\mathbf{I})$ に従う．設定 (a) の $\mathbf{x}$ と $\mathbf{x}''$ は同じ分布に従うので，$\mathbf{y}$ と $\mathbf{y}''$ を比較することが可能となる．

定理 1 では，クエリ事例 $\mathbf{x}$ の分布を固定した時，ハブの出現を減らすためには，データ分布の分散は小さい方が望ましいことが示されている．よって，ハブの出現を減らすという観点から，$\mathbf{y}$ よりも $\mathbf{y}''$ が望ましい．すなわち，設定 (a) よりも設定 (b) の方が，ハブが出現しづらいことが予想される．

最後に，設定 (b) は本稿の提案法である，ラベルを事例空間に写像した状況をモデル化している．

4.3 事例とラベルの近傍関係に注目した分析

$\|\mathbf{z}\|$ に従って減少する確率密度関数 $p(\mathbf{z})$ の単峰分布を考えた場合，次の定理を得る．

定理 3 ユークリッド空間に存在する点の有限集合 Y を考える．各点は，確率密度関数 $p(\mathbf{z})$ が，$\|\mathbf{z}\|$ の減少関数となっている分布から独立にサンプルされるとする．任意の点 $\mathbf{y} \in Y$ と定数 $r > 0$ を定める．さらに，$\mathbf{x}_1$ と $\mathbf{x}_2$ を $\mathbf{y}$ から距離 r の点とする．この時，$\|\mathbf{x}_1\| < \|\mathbf{x}_2\|$ ならば，$\mathbf{y}$ が $\mathbf{x}_2$ の最近傍点になる確率は $\mathbf{x}_1$ の最近傍点なる確率よりも高い．

図 2 に定理 3 のイメージを示す．既存法は，事例 $\mathbf{x}$ を写像するので，ラベル $\mathbf{y}$ よりもノルムが小さくなる傾向にある．一方，提案法は，ラベル $\mathbf{y}$ を写像するので，事例 $\mathbf{y}$ よりもノルムが小さくなる傾向にある．

定理 3 は，事例 $\mathbf{x}$ の最近傍点をラベル $\mathbf{y}$ としたいならば，$\mathbf{y}$ のノルムを $\mathbf{x}$ のノルムよりも小さくする方が良いことを示唆している．よって，提案法を用いる方が各ラベルが最近傍になる確率が高くなることが期待できる．繰り返しになるが，この定理はハブ現象とは直接関連がないものの，提案法の正当性を補完する．

4.4 提案法のまとめ

4.1–4.3 節の分析より，本稿では，ラベルを事例空間へ写像し，事例空間において近傍検索を行うことを提案する．この方法は既存法（回帰を用いた ZSL）[7], [8], [13], [17], [19] とは逆方向の写像となる．

提案手法では，定理 2 における行列 $\mathbf{B}$ は事例 $\mathbf{X}$ を表してお

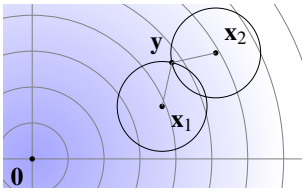

図 2 定理 3 の概略図．ここでは，$\|\mathbf{x}_1\| < \|\mathbf{x}_2\|$ と $\|\mathbf{y}-\mathbf{x}_1\| = \|\mathbf{y}-\mathbf{x}_2\|$ を仮定している．背景の明暗の濃さは Y の確率密度関数の値を表している．$\mathbf{x}_1$ を中心とした円の確率密度は，$\mathbf{x}_2$ を中心とした円の確率密度よりも大きい．

り，$\mathbf{A}$ はラベル $\mathbf{Y}$ を表している．すなわち，$\|\mathbf{MA}\|_2 \leq \|\mathbf{B}\|_2$ は $\|\mathbf{MY}\|_2 \leq \|\mathbf{X}\|_2$ となり，結果として，射影されたラベルのノルムは事例のノルムよりも小さくなりやすい．

本節の分析はデータの分布に強い仮定（例えば正規分布）を用いているため，現実には実データに適用できない．しかしながら，6 節において提案手法の有効性を実データを用いて確認している．

5. 関連研究

Palatucci ら [19] は，初めて ZSL の写像関数としてリッジ回帰を用いた．その後，今日までリッジ回帰は ZSL における標準的なアプローチの一つとなっている．特に自然言語処理においては，句生成 [7]，対訳抽出 [7], [8], [17] で使われている．近年，非線形写像を行うためにニューラルネットワークを写像関数として用いる研究も行われている [9], [22]．これら全ての回帰ベースの手法は，事例をラベル空間に写像する関数を学習している．

ZSL は正準相関分析の問題としても定式化できる．Hardoon ら [10] は正準相関分析とカーネル正準相関分析を画像ラベル付けに用いた．Lazaridou ら [13] は画像ラベル付けにおいて，リッジ回帰，正準相関分析，特異値分解，ニューラルネットワークの比較を行った．本稿の実験でも（6 節），比較のため正準相関分析をベースラインの一つとして用いた．

Dinu and Baroni [8] はハブの出現が ZSL の予測精度に悪影響を与えていることを初めて報告した．彼らは，ハブの出現を抑制するために，ZSL の近傍探索に類似度の再計算を提案した．しかし，この手法はコサイン類似度にのみ適用できるものであり，距離に用いることはできない．

ZSL と同様の問題設定として structured output learning [4] がある．ただし，ZSL とは異なり structured output learning はラベルが複雑な構造を持っていることを仮定しており，結果として，構造データを特徴空間に埋め込む計算コストが大きい．Kernel dependency estimation [25] は，カーネル化した主成分分析とリッジ回帰を用いることで，この問題の解消を目指している．これによって，ラベル空間における近傍検索はカーネル空間における pre-image 問題 [15] となるが，これ自体も困難な問題である．

6. 実験

本実験では，人工データ・実データの双方で提案手法の評価を行う．本実験の目的は，4 節で述べた通り提案手法がハブの出現を抑制し，精度の改善が見られるかどうかを検証することである．

6.1 実験設定

6.1.1 比較手法

本実験では以下の手法を評価する．

- $\text{Ridge}_{\text{X}\to\text{Y}}$: 線形リッジ回帰．事例 X をラベル空間へ写像する．これは先行研究で用いられたものである [7], [8], [13], [17], [19].
- $\text{Ridge}_{\text{Y}\to\text{X}}$: 線形リッジ回帰．ラベル Y を事例空間へ写像する．4.4 節で述べたように，これが本論文の提案手法である．
- CCA: 正準相関分析 [10]．本実験では公開されているコード [*1] を使用した．

リッジ回帰に用いる正則化パラメータと CCA における次元数は，訓練セットの交差検定により決定した．

リッジ回帰（もしくは CCA）によって写像後，X と Y は同じ空間に存在するため，与えられた事例に対して，ユークリッド距離によってラベルの近傍検索を行う．加えて，単純なユークリッド距離ではなく，*non-iterative contextual dissimilarity measure* (NICDM) [11] による近傍検索も行う．NICDM はユークリッド距離によって作られた近傍関係をより対象になるように距離の再計算を行う．この尺度は，ZSL 以外のタスクにおいてハブが削減されることが報告されている [21].

なお，本実験で用いる全てのデータは，前処理として中心化が行われているものとする．

6.1.2 評価指標

本実験では (i) ZSL の予測精度と (ii) 近傍検索におけるハブの出現度合いの 2 種類の観点から評価を行う．

6.1.2.1 予測精度

本実験では，ZSL を検索問題として定式化する．すなわち，事例が与えられ，その事例に最も関係するラベルをラ

[*1] http://www.davidroihardoon.com/Professional/Code.html

ベル集合から検索する問題となる．検索問題として定式化したので，評価指標に mean average precision (MAP) [14] を用いる．注意点として，人工データと画像ラベル付け実験は事例に対応するラベルは 1 個のみしか存在しない．その場合，MAP は mean reciprocal rank [14] と同等の値を得る．MAP に加えて，上位 k 番目までの精度 [*2] (Acc_k) も示す．

6.1.2.2 ハブの出現度合い

先行研究 [20], [21], [23], [24] に従い，ハブの出現度合いを調べる指標として N_k 分布の歪度（N_k skewness）を用いる．N_k 分布は，全ての評価事例に対して評価ラベル i が上位 k 番目までに何回含まれたかを要素（$N_k(i)$）に持つ分布であり，その歪度は

$$(N_k \text{ skewness}) = \frac{\sum_{i=1}^{\ell} \left(N_k(i) - \mathrm{E}\left[N_k\right]\right)^3 / \ell}{\mathrm{Var}\left[N_k\right]^{\frac{3}{2}}}$$

によって求まる．ここで，ℓ は評価ラベルの数である．N_k 歪度が大きい値は評価事例の k 近傍に頻繁に含まれるラベルが存在していることを意味しており，すなわち，ラベルにハブが出現していることを指し示す．

6.2 タスクとデータセット

以下のタスクで評価を行う．

6.2.1 人工タスク

ZSL のタスクの模擬実験を行うために，それぞれが異なる空間に存在する事例とラベルのペアを生成する．まず，3000 次元のベクトル $\{\mathbf{z}_i \in \mathbb{R}^{3000} \mid i = 1, \ldots, 10000\}$ を生成する．ここで，各次元は独立で，同じ標準正規分布から生成されることを仮定している．この $\mathbf{z}_i$ を潜在ベクトルとする．すなわち，$\mathbf{z}_i$ を直接観測することはできないが，それに関連する事例 $\mathbf{x}_i$ とラベル $\mathbf{y}_i$ を観測することができる．この事例とラベルのペアは $\mathbf{x}_i = \mathbf{R}_X \mathbf{z}_i$ と $\mathbf{y}_i = \mathbf{R}_Y \mathbf{z}_i$ により生成される．この $\mathbf{R}_X, \mathbf{R}_Y \in \mathbb{R}^{300 \times 3000}$ はランダム行列であり，行列の各要素は区間 $[-1, 1]$ 上の一様分布からサンプルした．ランダムプロジェクションは写像前の空間の距離や角度を高確率で保存するので [5], [6]，写像されたオブジェクト集合は，異なった空間に存在しているが，類似した特徴を持っていることが期待できる．

最後に，生成した事例とラベルのペアの集合 $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{10000}$ をランダムに訓練セット（8000 ペア）と評価セット（2000 ペア）に分割した．

6.2.2 対訳抽出

一つ目の ZSL の実タスクは対訳抽出 [7], [8], [17] である．このタスクは，原言語の単語が与えられた際に，目的言語の対訳単語を非対訳単語よりも高く順位付けすることを目的としている．

本実験では，英語を目的言語として，以下の言語を原言語とした: Czech (cs), German (de), French (fr), Russian (ru), Japanese (ja), and Hindi (hi)．従って，本稿における ZSL の定式化では，各 6 言語を X とし，英語を Y とする．[*3]

先行研究 [7], [8], [17] に従い，Polyglot project[*4] [3] が配布している前処理済みの Wikipedia と word2vec[*5] を用いて単語ベクトルの学習を行った．学習に際して CBOW モデル [16] を選択し，ウィンドウサイズは 10，次元数は 500 とした．

写像関数の学習と予測精度の評価のために，Ács ら [1] が配布している対訳辞書 [*6] をゴールドの対訳対として用いた．このゴールドをランダムに訓練セット（80%）と評価セット（残りの 20%）に分割した．この分割を 4 回繰り返し，各評価の平均を最終的な実験結果とした．

6.2.3 画像ラベル付け

2 個目の実タスクは画像ラベル付けである．このタスクは，画像が与えられた際に，適した単語ラベルを検索する問題である．従って，ZSL の定式化では X がイメージであり，Y が単語ラベルとなる．

本実験では，50 クラス，30,475 画像からなる Animal with Attributes（AwA）データセット [*7] を用いる．画像の特徴ベクトルは，畳み込みニューラルネットワークによって学習された 4096 次元のベクトルを用いた．この特徴ベクトルは DeCAF と呼ばれており，AwA のウェブサイトで配布されている．計算量を削減するために，画像の特徴ベクトルをランダムプロジェクションにより，500 次元に削減した．

単語ラベルのベクトルは対訳抽出と同様に word2vec を用いて学習した．ただし，このタスクでは AwA のクラスラベル全ての単語ベクトルを構築するために，（2015 年 3 月 4 日時点の）Wikipedia を用いて学習した．word2vec のパラメータは対訳抽出と同じパラメータに設定した．

AwA で標準的に用いられる ZSL の設定に従い，データセットを訓練セット（40 ラベル）と評価セット（10 ラベル）に分割した．

6.3 実験結果

表 1 に実験結果を示す．実験結果の傾向は明快で，全てのタスクで提案手法 $\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ が MAP，Acc_k ともに他の手法を上回る結果となった．また，$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ と CCA は NICDM を用いた場合の方がユークリッド距離を用いる

[*2] 画像ラベル付け実験では，最近傍検索のマクロ平均（Acc_1）を報告している．

[*3] 各 6 言語を Y とし，英語を X とした場合の実験も行ったが，本実験と同様の結果を得た．

[*4] https://sites.google.com/site/rmyeid/projects/polyglot

[*5] https://code.google.com/p/word2vec/

[*6] http://hlt.sztaki.hu/resources/dict/bylangpair/wiktionary_2013july/

[*7] http://attributes.kyb.tuebingen.mpg.de/

表 1 実験結果: MAP は mean average precision を示している．$\mathrm{Acc_k}$ は上位 k 番目までの正解率を示している．N_k は N_k 分布の歪度であり，大きい値はハブが出現していることを意味している（小さい値ほどハブが出現しておらず良い結果となっている）．各評価指標で最も良い数値をボールド体で示している．

(a) 人工データ．

method	MAP	Acc_1	Acc_{10}	N_1	N_{10}
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$	21.5	13.8	36.3	24.19	12.75
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ + NICDM	58.2	47.6	78.4	13.71	7.94
$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ (proposed)	**91.7**	**87.6**	**98.3**	**0.46**	**1.18**
CCA	78.9	71.6	91.7	12.0	7.56
CCA + NICDM	87.6	82.3	96.5	0.96	2.58

(b) 対訳抽出の MAP

method	cs	de	fr	ru	ja	hi
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$	1.7	1.0	0.7	0.5	0.9	5.3
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ + NICDM	11.3	7.1	5.9	3.8	10.2	21.4
$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ (proposed)	**40.8**	**30.3**	**46.5**	**31.1**	**42.0**	**40.6**
CCA	24.0	18.1	33.7	21.2	27.3	11.8
CCA + NICDM	30.1	23.4	39.7	26.7	35.3	19.3

(c) 対訳抽出の Acc_k．

	cs		de		fr		ru		ja		hi	
method	Acc_1	Acc_{10}	Acc_1	Acc_{10}	Acc_1	Acc_{10}	Acc_1	Acc_{10}	Acc_1	Acc_{10}	Acc_1	Acc_{10}
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$	0.7	2.8	0.4	1.6	0.3	1.2	0.2	0.8	0.2	1.3	2.9	8.2
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ + NICDM	7.2	17.9	4.3	11.4	3.5	9.8	2.1	6.3	6.1	16.8	14.4	32.6
$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ (proposed)	**31.5**	**54.5**	**21.6**	**43.0**	**36.6**	**58.6**	**21.9**	**43.6**	**31.9**	**56.3**	**31.1**	**55.4**
CCA	17.9	32.7	12.9	25.2	27.0	41.7	15.2	28.8	20.2	37.3	7.4	18.9
CCA + NICDM	21.9	42.3	16.1	33.9	31.1	50.1	18.7	37.0	25.9	48.8	12.4	30.7

(d) 対訳抽出の N_k．

	cs		de		fr		ru		ja		hi	
method	N_1	N_{10}	N_1	N_{10}	N_1	N_{10}	N_1	N_{10}	N_1	N_{10}	N_1	N_{10}
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$	50.29	23.84	43.00	24.37	67.79	35.83	95.05	35.36	62.12	22.78	23.75	10.84
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ + NICDM	41.56	20.38	39.32	20.82	57.18	25.97	89.08	30.70	57.57	21.62	20.33	9.21
$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ (proposed)	**11.91**	**10.74**	**12.49**	**11.94**	**2.56**	**2.77**	**4.28**	**4.18**	**5.15**	**6.76**	**10.45**	**6.14**
CCA	28.00	18.67	36.66	18.98	30.18	15.95	51.92	21.60	37.73	18.27	22.31	8.95
CCA + NICDM	25.00	17.13	32.94	17.65	25.20	14.65	42.61	20.72	34.66	13.16	22.00	8.46

(e) 画像ラベル付け．

method	MAP	Acc_1	N_1
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$	46.0	22.6	2.61
$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ + NICDM	54.2	34.5	2.17
$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ (proposed)	**62.5**	**41.3**	**0.08**
CCA	26.1	9.2	2.00
CCA + NICDM	26.9	9.3	2.42

場合と比べて良い結果を得ることがわかった．

N_k 分布の歪度に注目すると，$\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ が最も低い値を得ていることが確認できる．これはハブが削減できていることを意味する．対照的に，$\mathrm{Ridge}_{\mathrm{X}\to\mathrm{Y}}$ は高い値を得ており，予測精度に即した結果となっている．これらの結果は，4 節で議論した通りの結果となった．

また，ハブの出現度合いと ZSL の予測精度が逆相関にあることが確認できる．従って，ハブは予測精度に影響を与える重要な要素の一つであることがわかる．

AwA において，Akata ら [2] は畳み込みニューラルネットワークで学習したベクトルと word2vec で学習した 100 次元のベクトルを用いて 39.7%（Acc_1）精度を得たことを報告している．この結果と比較するために，提案手法 $\mathrm{Ridge}_{\mathrm{Y}\to\mathrm{X}}$ を同様な実験設定（DeCAF を次元圧縮せずに使い，word2vec で 100 次元のベクトルを再構築した）の下で評価した．結果として，提案手法は 40.0%（Acc_1）の結果を得た．この実験設定は厳密に同じ設定ではないので，直接比較することはできないが，説明変数と目的変数を入れ替えただけの素朴なリッジ回帰でも，Akata らが提案した洗練された手法と同等の性能を得る可能性があることが示唆できた．

7. まとめ

本稿では，ZSL における一般的な写像とは逆方向となる，ラベル空間から事例空間への写像について議論した．データが多変量正規分布に従うという単純なモデルを仮定し，ハブの出現という観点から，なぜ提案する写像方向が望ましいかについての説明を与えた．本実験の結果，提案手法はベースラインよりも，良い結果を得ることができた．

今後の予定として，(i) 4 節の議論を拡張すること，(ii) ニューラルネットワークなどのリッジ回帰以外の写像関数を用いた場合の影響を調査すること，(iii) CCA におけるハブの出現に関する分析を行うことが挙げられる．

参考文献

[1] Ács, J., Pajkossy, K. and Kornai, A.: Building basic vocabulary across 40 languages, *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, pp. 52–58 (2013).
[2] Akata, Z., Lee, H. and Schiele, B.: Zero-shot learning with structured embeddings, *arXiv preprint arXiv:1409.8403v1* (2014).
[3] Al-Rfou, R., Perozzi, B. and Skiena, S.: Polyglot: Distributed word representations for multilingual NLP, *CoNLL '13*, pp. 183–192 (2013).
[4] Bakir, G., Hofmann, T., Schölkopf, B., Smola, A. J., Taskar, B. and Vishwanathan, S. V. N.(eds.): *Predicting Structured Data*, MIT press (2007).
[5] Bingham, E. and Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data, *KDD '01*, pp. 245–250 (2001).
[6] Dasgupta, S.: Experiments with random projection, *UAI '00*, pp. 143–151 (2000).
[7] Dinu, G. and Baroni, M.: How to make words with vectors: Phrase generation in distributional semantics, *ACL '14*, pp. 624–633 (2014).
[8] Dinu, G. and Baroni, M.: Improving zero-shot learning by mitigating the hubness problem, *Workshop at ICLR '15* (2015).
[9] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ronzato, M. and Mikolov, T.: Devise: A deep visual-semantic embedding model, *NIPS '13*, pp. 2121–2129 (2013).
[10] Hardoon, D. R., Szedmak, S. and Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods, *Neural Computation*, Vol. 16, pp. 2639–2664 (online), DOI: 10.1162/0899766042321814 (2004).
[11] Jegou, H., Harzallah, H. and Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search, *CVPR '07*, pp. 1–8 (online), DOI: 10.1109/CVPR.2007.382970 (2007).
[12] Larochelle, H., Erhan, D. and Bengio, Y.: Zero-data learning of new tasks, *AAAI '08*, pp. 646–651 (2008).
[13] Lazaridou, A., Bruni, E. and Baroni, M.: Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world, *ACL '14*, pp. 1403–1414 (2014).
[14] Manning, C. D., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).
[15] Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M. and Rätsch, G.: Kernel PCA and de-noising in feature space, *NIPS '98*, pp. 536–542 (1998).
[16] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *Workshop at ICLR '13* (2013).
[17] Mikolov, T., Le, Q. V. and Sutskever, I.: Exploiting similarities among languages for machine translation, *arXiv preprint arXiv:1309.4168* (2013).
[18] 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, *ICLR '14* (2014).
[19] Palatucci, M., Pomerleau, D., Hinton, G. and Mitchell, T. M.: Zero-shot learning with semantic output codes, *NIPS '09*, pp. 1410–1418 (2009).
[20] Radovanović, M., Nanopoulos, A. and Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research*, Vol. 11, pp. 2487–2531 (2010).
[21] Schnitzer, D., Flexer, A., Schedl, M. and Widmer, G.: Local and global scaling reduce hubs in space, *Journal of Machine Learning Research*, Vol. 13, pp. 2871–2902 (2012).
[22] Socher, R., Ganjoo, M., Manning, C. D. and Ng, A. Y.: Zero-shot learning through cross-modal transfer, *NIPS '13*, pp. 935–943 (2013).
[23] Suzuki, I., Hara, K., Shimbo, M., Saerens, M. and Fukumizu, K.: Centering similarity measures to reduce hubs, *EMNLP '13*, pp. 613–623 (2013).
[24] Tomašev, N., Rupnik, J. and Mladenić, D.: The role of hubs in cross-lingual supervised document retrieval, *PAKDD '13*, pp. 185–196 (2013).
[25] Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A. and Schölkopf, B.: Kernel dependency estimation, *NIPS '02*, pp. 873–880 (2002).