Vol.2017-NL-231 No.23
Vol.2017-SLP-116 No.23
2017/5/16

# ソーシャルメディアにおける発言位置の分布表現とそれによる不確実性の推定

磯 颯[1,a)]　若宮 翔子[1]　荒牧 英治[1]

**概要**：ソーシャルメディアにおける発言位置の推定において，各単一の発言ごとの位置情報の推定 (Message level geolocation) に用いることの出来る情報には限りがあり，その不確実性を適切に評価することが必要である．本研究では，Convolutional Neural Network と Mixture Density Network を用いた End-to-end での条件付き正規混合分布の推定手法を提案し，Twitter 上の単一の発言に対し，その不確実性を考慮した位置情報の推定を行った．提案手法は，既存のアプローチである Classification による位置情報の推定より推定精度の意味では劣るものの，複数の指標において適切にその不確実性を評価出来ていることが分かり，提案手法の有用性を示した．

**キーワード**：位置情報，密度推定，ソーシャルメディア，Mixture Density Network，Convolutional Neural Network

## 1. はじめに

Twitter をはじめとするソーシャルメディアにおいて，その位置情報は感染症 [1] や選挙動向の把握 [2]，地震の早期検知 [3] など様々な領域で応用されている．一方，全ての発言に対して位置情報が付与されているわけではないため，その発言内容やユーザの情報などを用いてその発言位置を推定する必要がある．

発言位置の推定は大きく分けて User level と Message level の 2 種類の推定対象が存在する．前者の User level はユーザの居住地推定を行うもので選挙の予測や広告の配信などユーザプロファイルに依存したサービスに必要となる．一方，後者の Message level は各発言に対しその発言位置の推定を行うもので混雑状況の推定や店舗の評判など移動経路に依存したサービスに用いることができる．User level の位置情報を推定する場合，一般にユーザの発言が増加するに伴い，その特定が容易となるが，Message level の場合，扱うことの出来る情報はたかだか一回の発言であり，人間であってもその特定が困難な発言が多く存在する．

本研究では，これまで多く用いられていた分類アプローチ（以降，Classification と呼ぶ）でなく，ニューラルネットワークを用いた密度推定により，発言内容やユーザのプロファイルなどの情報で条件付けられた推定位置情報を条件付き分布の推定を行う．これにより，推定発言位置の不確実性を分布の広がりとして解釈することが可能となり，定量的な推定位置情報の不確実性が評価可能となる．

テキストデータから，直接その条件付き分布の密度推定を行うモデルとして，Convolutional Neural Network [4,5] と Mixture Density Network [6] を組み合わせることで，End-to-End での条件付き正規混合分布の推定を行う Convolutional Mixture Density Network (CMDN) を提案する．

Convolutional Mixture Density Network (CMDN) を Shared task（W-NUT Geolocation Prediction in Twitter [7]）を用いた実験では，位置座標の推定精度の意味では Classification による手法に劣るものの，複数の指標において適切にその不確実性を評価出来ていることが分かり，提案手法の有用性を示した..

なお，本研究におけるデータならびにコードは Web 上に公開している[*1]．

## 2. 関連研究

ソーシャルメディアにおける位置情報の推定は，Facebook [8] や Wikipedia [9], Flickr [10] など，様々なプラットフォームで行われている．特に Twitter を対象とした位置特定を目指す研究は，そのデータの適用範囲の広さ

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] iso.hayate.id3@is.naist.jp

[*1] http://sociocom.jp/~iso/geoden/

© 2017 Information Processing Society of Japan

から最も注目されているといえる [11]．Twitter における位置情報の推定を行う研究は，推定対象，用いる特徴量の種類，出力形式の 3 つの観点から大別することができる．

### 推定対象

1 章でも述べたように，Twitter における位置情報の推定には，User level と Message level の大きく 2 種類の推定対象が存在する．このうち，より詳細な情報を扱うことのできる Message level の位置情報がより実応用に際し重要であると言われている [7]．その一方で，多くの先行研究は User level での位置情報の推定にとどまっており [12–21]，Message level での位置情報を扱う研究は少数である [22–24]．

### 効果的な特徴抽出手法

推定精度向上のため，様々な有用な特徴量を用いたモデルが考案されてきた．Han et al. (2012) [17] は，対応する位置固有の発言を抽出を用いて発言位置の推定を行った．Jurgens et al. (2015) [19] や Rahimi et al.(2015) [20] は，関連の高いユーザ同士が類似の地域に居住していることを仮定し，ユーザの follower-followee 関係を用いて居住地の推定を行っている．Liu and Huang (2016) [24] は，ユーザの遷移情報を用いることで，Message level の位置情報推定の精度を大きく向上させた．Dredze et al. (2016) [23] は，発言の時間や曜日の情報が，Message level の位置情報の特定に効果的であることを示した．また，近年開催された W-NUT Twitter geolocation prediction shared task [7] において，User level, Message level いずれにおいても距離の誤差の意味で最も高い精度を示した Miura et al. (2016) [25] は，ツイートの記述のみでなく，ユーザの自由記述による個人情報開示欄や居住地欄，また，各ユーザや発言のタイムゾーンなどのユーザのメタ情報が推定に大きく寄与することを示した．

これらの特徴抽出手法のうち，ユーザのネットワーク関係や遷移情報を用いる場合，推定精度の向上が期待できる一方で，コーパス中に同じネットワークに属するユーザの存在や，各ユーザの発言数に大きく依存する．そのため本研究では，単一の発言に紐づく情報のみから推定が可能かつその中で最も精度の高い Miura et al. (2016) [25] と同様にテキストデータとメタデータそれぞれから独立に特徴抽出を行いそれらの特徴を組み合わせる方式を用いる．

### 出力形式

特徴量から GPS 情報を直接推定する手法は少数であり，我々の調べにおいては，Eisenstein et al. (2010) [13] を筆頭とした生成モデルによるもののみとなっている [14,18]．

他の多くの研究は，学習データにおける GPS 情報を対応するクラスへ予め変換し，Classification の問題として定式化するアプローチが最も主流である．変換対象となるクラスとして，行政地方区分の県や都市レベルのクラスだけでなく，等間隔のグリッドに分割するもの (Wing et al. (2011) [15]) や，データの密集している領域ではより細かいグリッド，データの少ない領域では大きなグリッドを $k$-d tree を用いて分割するもの (Roller et al. (2012) [16]) などが提案されている．

また，推定精度を向上させるためには，グリッドサイズを小さくする必要があるが，グリッドのサイズが小さくなるごとに，各グリッドの占めるデータ量が小さくなってしまうデータスパースネスの問題のトレード・オフが存在する．そのため，Hulden et al. (2015) [21] は，各グリッドでのカウントを用いるのではなく，ガウス窓関数を用いて周辺領域のグリッドにおけるカウントも考慮することでデータスパースネスの問題を解消し，同時に位置情報の推定精度向上を達成している．

最後に，本研究と最も関連の深い Priedhorsky et al. (2014) [22] では，各発言やメタデータにおける $n$-gram それぞれにおいて正規混合分布を推定し，その推定分布の重み付け和により Message level での密度推定を行い，推定分布からの測定を行っている．Priedhorsky et al. (2014) [22] による手法の問題点として，各発言を独立に混合正規分布やその重要度を推定するため，各発言に対する分布の推定に 2 step 必要であるだけでなく，過剰に重要度を高く見積もってしまうことが，我々の実験でわかった (see 表 4)．これに対し，本研究で用いるニューラルネットワークによる条件付き混合分布の推定は，特徴抽出から分布のパラメータ推定まで End-to-end で行うことができる．

## 3. Model

本章では，提案手法である，Convolutional Mixture Density Network (CMDN) について述べる．これは，Bishop (1994) [6] により提案された Mixture Density Network (MDN) の特徴抽出部を Kim (2014) [26] による Convolutional neural network を用いた文書分類モデルに置き換え，よりテキストデータに適した形に拡張したものに相当する．

### 特徴抽出

まず，ツイートの長さ $L$，$w_i \in \mathbb{R}^V$ を $i$ 番目の単語を表す one-hot ベクトルとし，各ツイートを $\mathbf{w} = \mathbf{w}_1 \oplus \mathbf{w}_2 \oplus \cdots \oplus \mathbf{w}_L$ として表す．ここで，$\oplus$ はベクトルの連結を表す作用素とする．

各単語ベクトル $\mathbf{w}_i$ に対し，その低次元表現 $\mathbf{x}_i$ を射影行列 $\mathbf{W}_e$ により，$\mathbf{x}_i = \mathbf{W}_e\mathbf{w}_i \in \mathbb{R}^d$ と表す．ここで，$V$ は語彙数，$d$ は分散表現の次元を表す．これより，ツイートの文章ベクトル $\mathbf{x}_{1:L}$ を，単語ベクトル $\mathbf{x}_i$ を連結するこ

© 2017 Information Processing Society of Japan

とで以下のように示す．

$$\mathbf{x}_{1:L} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \cdots \oplus \mathbf{x}_L \in \mathbb{R}^{Ld}.$$

この文章ベクトルから有用な特徴を取り出すため，フィルタ行列 $\mathbf{W}_f$ を各文章ベクトルに対し窓幅 $\ell$ で適用する．

$$\mathbf{W}_f = [\mathbf{w}_{f,1}, \mathbf{w}_{f,2}, \ldots, \mathbf{w}_{f,m}]^\top$$
$$c_{i,j} = \phi(\mathbf{w}_{f,j}^\top \mathbf{x}_{i:i+\ell} + b_{f,j})$$

ここで，$m$ をフィルタ数，$\phi$ を活性化関数とする．本研究では，活性化関数として，ReLU [27] を用いる．

各フィルタの出力に対し，その最大値を各フィルタに対し最も応答の大きなユニットの出力を $\hat{c}_j$ として取り出す (1-max pooling)．これらの特徴量を連結し，以下のように特徴ベクトル $\mathbf{h}$ を構成する．

$$\hat{c}_j = \max_{i \in \{1,\ldots,L-\ell+1\}} c_{i,j}$$
$$\mathbf{h} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_m]^\top$$

### ニューラルネットワークによるパラメータ推定

Kim (2014) [26] による文書分類モデルの場合，得られた特徴ベクトル $\mathbf{h}$ を用いて Classification モデルを構築していた．MDN を用いた条件付き分布 $\hat{p}(\mathbf{y} \mid \mathbf{w})$ の密度推定の場合，混合分布のパラメータを特徴量 $\mathbf{h}$ から推定する．特に本研究では，$K$ 個の $q$ 次元多変量正規分布 $\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ の組み合わせにより表現される条件付き正規混合分布

$$\hat{p}(\mathbf{y} \mid \mathbf{w}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

を仮定し，パラメータ $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ の推定を目的とする．

各 $q$ 次元多変量正規分布に対し，推定すべきパラメータ数 $p$ は，$p = \frac{q(q+3)}{2}$ 個となる．(平均 $\boldsymbol{\mu}_k$，分散共分散行列の対角成分に $\boldsymbol{\Sigma}_k$ に対しそれぞれ $q$ 個，分散共分散行列の相関項 $\rho$ に対し $\frac{q(q-1)}{2}$ 個)．

ここで $q = 2$ の場合，各混合分布を構成する 2 変量正規分布は以下のように表わされる．

$$\boldsymbol{\mu}_k = \begin{pmatrix} \mu_{k,1} \\ \mu_{k,2} \end{pmatrix}, \boldsymbol{\Sigma}_k = \begin{pmatrix} \sigma_{k,1}^2 & \rho_k \sigma_{k,1} \sigma_{k,2} \\ \rho_k \sigma_{k,1} \sigma_{k,2} & \sigma_{k,2}^2 \end{pmatrix}.$$

また，混合分布の混合比 $\pi_k$ として，$K$ 個のパラメータが必要となる．これらのパラメータを推定するために，特徴ベクトル $\mathbf{h}$ から混合分布の推定に必要なパラメータ数 $K(p+1)$ 個と同次元のベクトル $\boldsymbol{\theta}$ を以下のように構成する．

$$\boldsymbol{\theta} = \mathbf{W}_p \mathbf{h} + \mathbf{b}_p \in \mathbb{R}^{Kp},$$

ここで，$\mathbf{W}_p \in \mathbb{R}^{Kp \times m}$，$\mathbf{b}_p \in \mathbb{R}^{Kp}$ はそれぞれ係数行列と切片項を表す．

ここで得たベクトル $\boldsymbol{\theta}$ は，$K$ 個の q 次元混合正規分布のパラメータ数と同じだけ値を保持しているが，ベクトル $\boldsymbol{\theta}$ の各成分は実数値 $\mathbb{R}$ を保持しており，多変量正規分布のパラメータ $\sigma_k, \rho_k$ や，その混合比 $\pi_k$ の値として適切ではない．混合比 $\pi_k$ は常に正の値をとり，$k$ についての和が 1 である必要がある．また，分散 $\sigma$ は，正の実数である必要がある．相関項 $\rho_k$ は，$(-1, 1)$ の元である必要がある．このため，我々は実数値出力 $\boldsymbol{\theta}$ を，混合分布のパラメータとして適切な値を取るように値域を制限する必要がある．

### パラメータの変換

まず，$\boldsymbol{\theta}$ を，その構成分布ごとの推定パラメータ $\boldsymbol{\theta}_k$ として以下のように分解して考える．

$$\boldsymbol{\theta} = \boldsymbol{\theta}_1 \oplus \boldsymbol{\theta}_2 \oplus \cdots \oplus \boldsymbol{\theta}_K$$
$$\boldsymbol{\theta}_k = (\theta_{\pi_k}, \theta_{\mu_{k,1}}, \theta_{\mu_{k,2}}, \theta_{\sigma_{k,1}}, \theta_{\sigma_{k,2}}, \theta_{\rho_k}).$$

ここで，各パラメータの値域を制限するために，各 $\boldsymbol{\theta}_k$ に対し，以下の変換を適用する．

$$\begin{aligned}
\pi_k &= \mathrm{softmax}(\theta_{\pi_k}) \\
&= \frac{\exp(\theta_{\pi_k})}{\sum_{k'=1}^{K} \exp(\theta_{\pi'_k})} \in (0, 1) \\
\mu_{k,j} &= \theta_{\mu_{k,j}} \in \mathbb{R}, \\
\sigma_{k,j} &= \mathrm{softplus}(\theta_{\sigma_{k,j}}) \\
&= \ln\left(1 + \exp(\theta_{\sigma_{k,j}})\right) \in (0, \infty), \\
\rho_k &= \mathrm{softsign}(\theta_{\rho_k}) \\
&= \frac{\theta_{\rho_k}}{1 + |\theta_{\rho_k}|} \in (-1, 1).
\end{aligned}$$

これにより，各パラメータが多変量正規分布や，混合比のパラメータとして適切な値を取ることができる．

Mixture Density Network の原論文 [6] や，手書き文字生成への応用に Mixture Density Network を用いた研究 [28] では指数関数 exp を分散 $\sigma$ の変換に，双曲線正接関数 tanh を相関項 $\rho$ に適用していたが，我々は，softplus 関数 [29] を分散，softsign 関数を [30] 相関項へ適用した．これまで用いられていた指数関数や双曲線正接関数は，勾配消失や発散が起こり最適化が困難となる問題が生じていたが，値域の変換に softplus や softsign 関数を用いることでより頑健な推定が可能となる．

他にも最適化の際に必要となる数値計算上のテクニックに関しては，Appendix に付記する．

### パラメータ推定のための損失関数

混合分布の最適化に対し，損失関数 $L$ として負の対数尤度を最小化することで推定分布の最適化を行う．

© 2017 Information Processing Society of Japan

| WNUT Twitter Geolocation Prediction Dataset | | |
|---|---|---|
| Train | # of tweets | 9,339,618 |
| | # of users | 760,176 |
| | Avg. # of word in text | 9.51 |
| | # of vocabulary in text | 210,910 |
| | Avg. # of word in location | 1.66 |
| | # of vocabulary in location | 67,958 |
| | Avg. # of word in description | 8.66 |
| | # of vocabulary in description | 367,259 |
| Dev | # of tweets | 7,036 |
| | # of users | 4,582 |
| Test | # of tweets | 10,000 |
| | # of users | 6,434 |

表 1: データセットにおける統計量

| Fixed parameters | |
|---|---|
| Embedding dimension | 256 |
| Categorical embedding dimension | 64 |
| Window sizes | 3, 4, 5 |
| Each filter size | 128 |
| Batch size | 1500 |
| Learning rate | 0.001 |
| Optimization | Adam [31] |

表 2: 固定パラメータ

$$L = -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right).$$

## 4. 実験

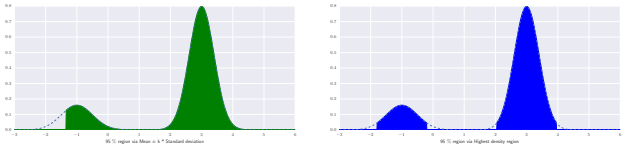
本研究では，W-NUT Geolocation Prediction in Twitter [7] のデータ (表 1) を対象に，密度推定により得られた推定分布からその発言の推定精度を検証するとともに，主に Lakshminarayanan et al. (2016) [33] に従い，推定分布における不確実性の評価を行う．

### 4.1 位置情報の推定精度の比較

各モデルにおいて，発言位置の推定精度の比較を行う．各発言の推定座標と実際の座標との距離は haversine 式[*2]を用いて計算する．評価尺度として，テストデータにおける距離誤差の中央値 (Median)，平均値 (Mean)，推定座標が，実際の発言位置から 161 km (100 miles) 以内に存在する確率 (Acc@161) により評価を行う．多くの場合，大きな距離誤差を多く含むため外れ値の影響を受けやすい平均値だけでなく中央値による評価も行っている．

なお，Classification による推定に対しては，その都市

[*2] https://en.wikipedia.org/wiki/Haversine_formula



(a) 平均 $\pm k\times$ 標準偏差　(b) Highest Density region

図 1: 混合正規分布における 95 % 信頼区間

レベルでの分類精度も付記する．

**推定分布による位置の推定**

密度推定により推定された分布からその推定発言位置として代表点を決める必要がある．本研究では，Mixture Density Network の原論文 [6] に従い，推定分布の最頻値を推定位置 $\hat{\mathbf{y}}$ として用いる．一般の確率分布に対し，数値的な最適化によりその最頻値を探索することが可能であるが，計算量が膨大であり大規模データに対しスケールさせることは難しい．そのため，近似的な最頻値として混合分布における各正規分布の平均値のうち，最も尤度が高くなる値，即ち

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}}{\operatorname{argmax}} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

として，得られた値 $\hat{\mathbf{y}}$ を条件付き分布における推定座標として用いる．この方法は，論文中に明示的に述べられていないものの，Priedhorsky et al. (2014) [22] による実装[*3]でも同じ方法で推定分布から近似的な最頻値を得ている．

### 4.2 推定分布の評価

推定された分布を用いて発言における位置情報の不確実性を適切に捉えることができているか評価する．例えば，推定分布において 20%の確率で存在すると推定された領域内にテストデータの 20%が存在していれば，推定分布は適切に不確実性を評価できていると考えられる．このような推定結果の評価方法は Reliability diagrams や Calibration curves として知られている．本研究では，1%から 99%まで 1%刻みで描いた Reliability diagrams を用いて評価する．

さらに，負の対数尤度を用いることでテストデータにおける推定分布の当てはまりのよさを評価する．負の対数尤度で評価することで，高い信頼度での推定が外した場合に負の対数尤度がより大きくなり，大きな誤りをした場合でも推定の信頼度が低いとモデルが評価した場合，負の対数尤度の値は小さくなる．

**多峰性のある分布に対する信頼域の推定**

不確実性の推定のため，各推定分布において，実際のテ

[*3] https://github.com/reidpr/quac/

© 2017 Information Processing Society of Japan

| | Method | Median (km) | Mean (km) | Acc@161 | City |
|---|---|---|---|---|---|
| Density estimation | CMDN-wide | 201.2 | 2009.1 | 0.420 | - |
| | MDN-wide | 532.7 | 2185.9 | 0.226 | - |
| | GMM-Err-SAE4 | 990.3 | 3491.2 | 0.319 | - |
| | GMM-Err-SAE10 | 1007.4 | 3519.2 | 0.317 | - |
| Classification | CNN-wide | **53.9** | **1788.9** | **0.576** | **0.437** |
| | MLP-wide | 67.4 | 1903.9 | 0.561 | 0.420 |
| | Miura et al. (2016) [25] | 69.5 | 1792.5 | - | 0.409 |
| | Jayasinghe et al. (2016) [32] | 74.7 | 2538.2 | - | 0.436 |
| Regression analysis | CNN-wide-l1 | 2927.6 | 4178.1 | 0.01 | - |
| | MLP-wide-l1 | 3588.2 | 4713.3 | 0.01 | - |
| | Major | 11718.4 | 10259. | 0.06 | 0.03 |

表 3: The prediction performance for WNUT message level geolocation.

| Method | Negatile log-likelihood | |
|---|---|---|
| | Mean | Median |
| CMDN-wide | $\mathbf{5.534 \pm 0.044}$ | $\mathbf{5.288 \pm 0.081}$ |
| MDN-wide | $6.713 \pm 0.039$ | $6.206 \pm 0.049$ |
| GMM-Err-SAE4 | $261.8 \times 10^6 \pm 184.9 \times 10^5$ | $10.199 \pm 0.133$ |
| GMM-Err-SAE10 | $261.6 \times 10^6 \pm 186.1 \times 10^5$ | $9.684 \pm 0.128$ |

表 4: Predictive Negative log-likelihood on test data

ストデータが何%の信頼域に含まれていたのか評価する必要がある．しかし，図 1a に示すように，正規混合分布のような多峰性のある分布に対し，平均と標準偏差を用いた信頼域の推定は適切ではない．

そこで，本研究では Highest density region (HDR) [34] を用いることで，実際のテストデータの座標は，推定分布において何%の信頼域に含まれていたのか評価する．Highest density region は，ある尤度以上の領域を推定域とする区間推定法であり，推定域の確率が事前に設定した信頼係数 1 - $\alpha$ となるように尤度を設定することで区間推定を行う．これにより，図 1b に示すように多峰性のある分布に対する信頼域を適切に評価することが出来る．

実際，Highest density region を用いたテストデータの含まれる信頼域の信頼係数の推定のため，あるツイート $\mathbf{w}_{obs}$ により条件づけられた推定分布　$\hat{p}(\mathbf{y}' \mid \mathbf{w} = \mathbf{w}_{obs})$ からサンプリングを行う．

$$\mathbf{y}'_1, \ldots, \mathbf{y}'_S \sim \hat{p}(\mathbf{y} \mid \mathbf{w} = \mathbf{w}_{obs})$$

そのサンプルと実際のテストデータにおける尤度を比較し，推定分布におけるテストデータの推定信頼係数を以下のように評価する．

$$\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}\left[\hat{p}(\mathbf{y}'_s|\mathbf{w} = \mathbf{w}_{obs}) \leq \hat{p}(\mathbf{y}_{true}|\mathbf{w} = \mathbf{w}_{obs})\right]$$

### 4.3 比較手法

本研究では，比較モデルとして密度推定，Classification，回帰分析をそれぞれ比較する．Miura et al. (2016) [25] に従い，特徴量としてテキストデータとメタデータとして，ユーザの自由記述による個人情報開示欄や居住地欄と発言されたタイムゾーンを各モデルにおいて共通で用いる．

**Density estimation**

**CMDN-wide**: Convolutional Neural Network による特徴抽出を行い，Mixture Density Network により条件付き分布の密度推定を行う (see 3 章).

**MDN-wide**: 通常の Mixture Density Network [6].

**GMM-Err-SAE**$\alpha$: Priedhorsky et al. (2014) [22] による $n$-gram ごとに推定された正規混合分布の重み付け和により，発言ごとの条件付き分布の推定を行う．本研究では $n = 2, \alpha = 4, 10$ として設定する．また，重み付けの手法として，各 $n$-gram における推定分布の最頻値と，学習データにおいて対応する $n$-gram の生起する発言位置との平均誤差の逆数に $\alpha$ 乗し，それらを標準化したものを重みとして扱う．詳細は Priedhorsky et al. (2014) [22] を参照していただきたい．

**Classification**

**CNN-wide**: Kim (2014) [26] による Convolutional Neural Network を用いた　Classification .

**MLP-wide**: Multi layer perceptron による Classification .

**Miura et al. (2016)**: Miura et al. (2016) による W-NUT shared task の結果.

**Jayasinghe et al. (2016)**: Jayasinghe et al. (2016) による W-NUT shared task の結果.

© 2017 Information Processing Society of Japan

**Regression analysis**

**CNN-wide-l1**: Convolutional Neural Network により特徴抽出を行い，回帰分析により直接緯度経度の推定を行う．本研究では損失関数として平均絶対誤差 (Mean Absolute Error, MAE) を用いる．

$$L_{abs} = -\frac{1}{N}\sum_{n=1}^{N} \mid \mathbf{y}_n - \mathbf{W}_r\mathbf{h} + \mathbf{b}_r \mid$$

where $\mathbf{W}_r \in \mathbb{R}^{m\times 2}, \mathbf{b}_r \in \mathbb{R}^2$．多次元出力の回帰分析において，単純に誤差最小化により回帰モデルを直接推定した場合，各次元の出力における相関構造を考慮することが出来ないが，CNN-wide-l1 は共通の潜在特徴量を用いて回帰モデルを推定するため，出力間の相関構造を考慮した回帰モデルを推定することが可能である．これは，Reduced Rank Regression [35] や Matrix regression [36]，Multitask 学習などと同様のアプローチである．

**MLP-wide-l1**: Multi layer perceptron により特徴抽出を行い，CNN-wide-l1 と同じ損失関数を用いる．

**Major**: 学習データにおいて最も多く生じるラベルを全てのデータに対する推定値として用いる．

### 4.4 パラメータの設定

ニューラルネットワークによる推定において，表 2 に示すパラメータを全て固定して実験を行う．CMDN-wide, MDN-wide において，$\{4, 8, 16, 32, 64\}$ のうち，開発セットにおける損失の最も小さくなる混合数 $K$ を用いる．また，Classification モデル CNN-wide, MLP-wide に対する正則化手法として Dropout [37] を用いる．Dropout 率を 0.5 として推定を行う．一方，位置座標を直接実数値として推定する密度推定，回帰分析手法において Dropout による正則化による精度の向上が見られなかった．そのため，正則化として開発セットにおける損失を最も小さくする early-stopping のみを用いた．

### 4.5 前処理

本研究では，Twokenizer[*4] [38] を用いて単語の分割を行い，Preprocessor[*5] を用いて，URL，ユーザ名，絵文字をそれぞれ *URL, MENTION, EMOJI* へ変換した．また，任意の数値を *NUM* へ変換した．

### 4.6 実験結果とその考察

#### 推定精度

表 3 に位置情報の推定精度を示す．Median, Mean, Acc@161 全てにおいて Classification モデルである CNN-wide が最も高い精度を示した．

一方，回帰分析を用いた推定は密度推定と比較しても大きく失敗していることが分かった．この原因として，回帰分析において仮定される等分散性が満たされないことや，類似の発言が複数の地点で観測されることによる平均への回帰 [39] と呼ばれる現象が生じていることがあげられる．

まず，回帰分析の推定において任意の特徴量によって条件付けられた出力の分散は全て同一であることが仮定されるが，特徴量ごとにその推定確度は大きくことなるため，仮定を大きく逸脱していることが分かる．また，Twitter 上の発言は類似の内容のものが複数地点で多く観測される．平均絶対誤差による回帰モデルの推定の際，推定モデルは条件付き中央値を推定することになるため，こうした複数候補が存在する場合，そのいずれでもない点を推定値として返してしまう．こうした現象は平均への回帰と言われており[*6]，推定が困難となっている．

しかし，密度推定により回帰分析による問題を解決し，回帰分析モデルに比べ大きく精度が向上していることが分かる．これは，条件付き分布に回帰モデルのように分布の仮定を強く置かず，より一般的に推定しているため，回帰モデルによる位置情報の推定による限界を解決していることが分かった．

#### 推定分布の評価

まず，図 2 に，密度推定モデルにおける Reliability diagram を示す．

各 $n$-gram の GMM の重み付け和による密度推定モデル GMM-Err-SAE$\alpha$ による推定は，低い確率における推定確度評価は適切に行われているが，推定確度がより高いときの実際の推定は大きく逸脱していることが分かる．これに対し， Mixture Density Network を用いて推定された分布は，低い確率において GMM の重み付け和によるものに比べ，過剰になっているものの，全体としては，特に CMDN-wide が最も理想的な Reliability diagram を描いており，推定分布が適切にその推定確度を評価することができていることが分かる．

また，テストデータにおける負の対数尤度を用いた推定分布の当てはまりの良さを評価を表 4 に示す．GMM-Err-SAE$\alpha$ により推定されたモデルは，学習データにおいて地理的に局所的に存在した $n$-gram に対し，過度に分散を小さく推定したり，その推定された正規混合分布に対して過度に大きな重みを乗せてしまい，その $n$-gram が大きく逸脱した場所で生起した場合，非常に大きな負の対数尤度の値が得られる．そのため，負の対数尤度の平均がこうした外れ値に対して大きな影響を受けてしまうため，表 4

[*4] https://github.com/myleott/ark-twokenize-py
[*5] https://github.com/s/preprocessor

[*6] 平均絶対誤差の場合，条件付き中央値の推定を行っているため，正確には平均へ回帰するわけではない．一般に回帰分析に用いられる，平均自乗誤差 (Mean squared error, MSE) を用いて回帰モデルの推定を行った場合，その推定値が条件付き平均の推定量となるため，平均への回帰と呼ばれている．

© 2017 Information Processing Society of Japan

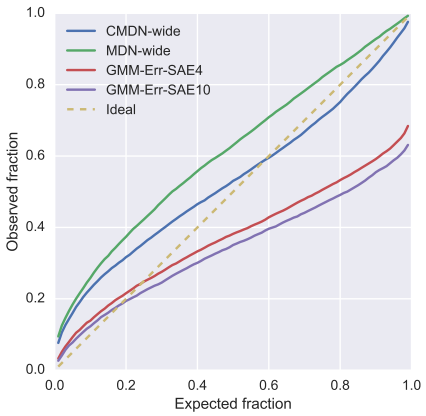



図 2: Reliability diagram: 推定分布における信頼域の確率とテストデータがその領域に含まれる割合の関係を表した図．例えば，推定分布の 50%信頼域に含まれるテストデータの割合が 50%に近ければ，適切にその不確実性を評価できているといえる．

において負の対数尤度の平均と中央値の両方を示した．

実際に GMM-Err-SAE$\alpha$ の負の対数尤度の平均値は，その中央値に比べ非常に大きな値をとっており，推定確度を過剰に高く評価する傾向があることが分かる．これは，図 2 の Reliability diagram において期待推定確度の高い発言に対し，適切にその推定確度を評価できなかったことの原因の一つと考えられる．．一方，Mixture Density Network による推定は負の対数尤度の平均，中央値のいずれの値も非常に近い値を示しており，推定確度を適切に推定できていることが分かる．

また，CMDN-wide による負の対数尤度の平均と中央値のいずれも最も小さな値を示しており，本研究で用いた密度推定モデルのうち最も高い汎化性能を示している．

## 5. 密度推定による位置情報推定の限界

本研究では，Mixture Density Network と，その特徴抽出に Convolutional Neural Network を用いて Message level のツイートの発言位置推定とその不確実性の評価を行った．我々の実験の結果，CMDN-wide による位置情報の推定は，直接位置座標の推定を行う手法の中で最も高い推定精度を示したが，Classification により推定されたラベルから代表点を選択し推定座標とする手法が最も高い推定精度を示した．この原因として，推定において Euclid 空間上で推定座標と実際の座標の誤差の評価を行っていることがあげられる．つまり，Euclid 空間上で同じ距離であっても，haversine 式を用いて測定する球面上の距離が大きくことなることがあげられる．

例えば，緯度経度が 0.0000°N, 0.0000°E から 0.0000°N, 80.0000°E までの Euclid 距離と，80.0000°N, 0.0000°E から 80.0000°N, 80.0000°E までの Euclid 距離は同じ 80 であるが，haversine 式を用いて図った場合，前者は約 8896 km，後者は約 1425 km 離れている．また，経度には周期性があり，0.0000°N, 180.0000°E から 0.0000°N, −180.0000°E は座標としては同一の点を指しているにも関わらず，Euclid 距離は 360 となってしまう．

このように，球面上の座標の値を直接 Euclid 空間上にマップし推定を行うことによる無視できない近似誤差が存在する．そのため，座標の値を Euclid 空間上の実数値として直接用いて，座標を直接推定する密度推定や回帰分析による推定には推定座標と実際の座標との距離誤差最小化の意味での推定精度には限界がある．

しかし，Convolutional Mixture Density Network による End-to-end での条件付き分布の密度推定による位置情報の推定は，Classification を除く手法の中で最も高い精度を示しているだけでなく，その推定の不確実性を最も適切に評価できていることがわかった．そのため，上記の球面上の点を 2 次元 Euclid 空間上の点として近似することによる誤差を解消することが出来れば，より位置情報の推定に適した推定手法となる可能性を秘めている．

## 6. まとめ

本研究では，End-to-end　推定が可能な条件付き分布の密度推定モデルとして Convolutional Mixture Density Network を提案し，Twitter の発言位置情報の推定と，その推定分布における不確実性の評価を行った．位置情報の推定精度は，球面上の点を 2 次元 Euclid 空間上の点として扱うことによる近似誤差により，既存の Classification によるアプローチに劣るものの，提案した密度推定モデルを用いた不確実性の評価指標として用意した Reliability diagram, 負の対数尤度のいずれにおいても最も高い精度でその不確実性を評価できることが分かった．

## 付　　録

### A.1 数値計算におけるテクニック

本章では，正規混合分布を推定する際，一般に用いられるテクニックと，Mixture Density Network による推定特有の数値計算におけるテクニックを紹介する．

まず，一般に正規混合分布における負の対数尤度を計算する際，数値計算における underflow を防ぐために log-sumexp と呼称されるテクニックを用いて数値計算の安定化を図る．logsumexp に関する詳細は高村 (2010) [40] を参考にされたい．

実際，一般の正規混合分布に対し負の対数尤度を計算する際，以下のように式変形をすることで，logsumexp による計算が出来る．

© 2017 Information Processing Society of Japan

$$
\begin{aligned}
L &= -\frac{1}{N}\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{y}_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\pi_k\frac{1}{\sqrt{(2\pi)^q\det\boldsymbol{\Sigma_k}}} \\
&\quad \exp\left(-\frac{1}{2}(\mathbf{y}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k)\right) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\ln\sum_{k=1}^{K}\exp\left(\ln\left(\frac{\pi_k}{\sqrt{(2\pi)^q\det\boldsymbol{\Sigma_k}}}\right)\right) \\
&\quad \exp\left(-\frac{1}{2}(\mathbf{y}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k)\right) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\mathrm{logsumexp}(z)
\end{aligned}
$$

where

$$
\begin{aligned}
z &= \ln\left(\frac{\pi_k}{\sqrt{(2\pi)^q\det\boldsymbol{\Sigma_k}}}\right)-\frac{1}{2}(\mathbf{y}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k) \\
&= \ln\pi_k-\frac{q}{2}\ln(2\pi)-\frac{1}{2}\det\boldsymbol{\Sigma}_k-\frac{1}{2}(\mathbf{y}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_n-\boldsymbol{\mu}_k)
\end{aligned}
$$

ここで，Mixture Density Network により得られるパラメータ $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ はサンプル $\mathbf{x}$ に依存し変動するため，特に混合数 $K$ を大きくした場合，いくつかの混合比 $\pi_k$ の値が 0 に近い値を取る．この際，$\ln\pi_k$ の値が発散してしまう．そこで，混合比 $\pi_k$ が 実数値を取る $\theta_{\pi_k}$ に対し，softmax 関数を用いて変換されていることから，

$$
\begin{aligned}
\ln\pi_k &= \ln\mathrm{softmax}(\theta_{\pi_k}) \\
&= \ln\left(\frac{\exp(\theta_k)}{\sum_{k=1}^{K}\exp(\theta_k)}\right) \\
&= \theta_k-\ln\sum_{k=1}^{K}\exp(\theta_k) \\
&= \theta_k-\mathrm{logsumexp}(\theta_k)
\end{aligned}
$$

として数値計算を行うことで，より安定した数値計算が可能となる．

## 参考文献

[1] Broniatowski, D. A., Paul, M. J. and Dredze, M.: National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic, *PLoS ONE*, Vol. 8, No. 12 (online), DOI: 10.1371/journal.pone.0083672 (2013).

[2] Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M. and Riotta, G.: A Multi-Level Geographical Study of Italian Political Elections from Twitter Data, *PLoS ONE*, Vol. 9, No. 5, pp. 1–11 (online), DOI: 10.1371/journal.pone.0095809 (2014).

[3] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, ACM, pp. 851–860 (online), DOI: 10.1145/1772690.1772777 (2010).

[4] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biological cybernetics*, Vol. 36, No. 4, pp. 193–202 (1980).

[5] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).

[6] Bishop, C.: Mixture Density Networks, Technical report (1994).

[7] Han, B., Hugo, A., Rahimi, A., Derczynski, L. and Baldwin, T.: Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text, *WNUT 2016*, p. 213 (2016).

[8] Backstrom, L., Sun, E. and Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 61–70 (2010).

[9] Lieberman, M. D. and Lin, J. J.: You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories., *ICWSM* (2009).

[10] Serdyukov, P., Murdock, V. and Van Zwol, R.: Placing flickr photos on a map, *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 484–491 (2009).

[11] Han, B., Cook, P. and Baldwin, T.: Text-based twitter user geolocation prediction, *Journal of Artificial Intelligence Research*, Vol. 49, pp. 451–500 (online), DOI: 10.1613/jair.4200 (2014).

[12] Cheng, Z., Caverlee, J. and Lee, K.: You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, New York, NY, USA, ACM, pp. 759–768 (online), DOI: 10.1145/1871437.1871535 (2010).

[13] Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P.: A Latent Variable Model for Geographic Lexical Variation, *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 1277–1287 (2010).

[14] Eisenstein, J., Ahmed, A. and Xing, E. P.: Sparse Additive Generative Models of Text, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1041–1048 (2011).

[15] Wing, B. P. and Baldridge, J.: Simple supervised document geolocation with geodesic grids, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Association for Computational Linguistics, pp. 955–964 (2011).

[16] Roller, S., Speriosu, M., Rallapalli, S., Wing, B. and Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, pp. 1500–1510 (2012).

[17] Han, B., Cook, P. and Baldwin, T.: Geolocation prediction in social media data by finding location indicative words, *Proceedings of COLING*, pp. 1045–1062 (2012).

[18] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J. and Tsioutsiouliklis, K.: Discovering geographical top-

© 2017 Information Processing Society of Japan

ics in the twitter stream, *Proceedings of the 21st international conference on World Wide Web*, ACM, pp. 769–778 (2012).

[19] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T. and Ruths, D.: Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice., *ICWSM*, pp. 188–197 (2015).

[20] Rahimi, A., Vu, D., Cohn, T. and Baldwin, T.: Exploiting Text and Network Context for Geolocation of Social Media Users, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, Association for Computational Linguistics, pp. 1362–1367 (online), available from ⟨http://www.aclweb.org/anthology/N15-1153⟩ (2015).

[21] Hulden, M., Silfverberg, M. and Francom, J.: Kernel Density Estimation for Text-Based Geolocation., *AAAI*, pp. 145–150 (2015).

[22] Priedhorsky, R., Culotta, A. and Del Valle, S. Y.: Inferring the origin locations of tweets with quantitative confidence, *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, pp. 1523–1536 (2014).

[23] Dredze, M., Osborne, M. and Kambadur, P.: Geolocation for Twitter: Timing Matters, *North American Chapter of the Association for Computational Linguistics (NAACL)* (2016).

[24] Liu, Z. and Huang, Y.: Where Are You Tweeting?: A Context and User Movement Based Approach, *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, New York, NY, USA, ACM, pp. 1949–1952 (online), DOI: 10.1145/2983323.2983881 (2016).

[25] Miura, Y., Taniguchi, M., Taniguchi, T. and Ohkuma, T.: A simple scalable neural networks based model for geolocation prediction in Twitter, *WNUT 2016*, Vol. 9026924, p. 235 (2016).

[26] Kim, Y.: Convolutional Neural Networks for Sentence Classification, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Association for Computational Linguistics, pp. 1746–1751 (online), available from ⟨http://www.aclweb.org/anthology/D14-1181⟩ (2014).

[27] Nair, V. and Hinton, G. E.: Rectified linear units improve restricted boltzmann machines, *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814 (2010).

[28] Graves, A.: Generating sequences with recurrent neural networks, *arXiv preprint arXiv:1308.0850* (2013).

[29] Glorot, X., Bordes, A. and Bengio, Y.: Deep Sparse Rectifier Neural Networks., *Aistats*, Vol. 15, No. 106, p. 275 (2011).

[30] Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks., *Aistats*, Vol. 9, pp. 249–256 (2010).

[31] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[32] Jayasinghe, G., Jin, B., Mchugh, J., Robinson, B. and Wan, S.: CSIRO Data61 at the WNUT geo shared task, *WNUT 2016*, p. 218 (2016).

[33] Lakshminarayanan, B., Pritzel, A. and Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, *arXiv preprint arXiv:1612.01474* (2016).

[34] Hyndman, R. J.: Computing and graphing highest density regions, *The American Statistician*, Vol. 50, No. 2, pp. 120–126 (1996).

[35] Izenman, A. J.: Reduced-rank regression for the multivariate linear model, *Journal of multivariate analysis*, Vol. 5, No. 2, pp. 248–264 (1975).

[36] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 69, No. 3, pp. 329–346 (2007).

[37] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting., *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958 (2014).

[38] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N. A.: Part-of-speech tagging for twitter: Annotation, features, and experiments, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, pp. 42–47 (2011).

[39] Stigler, S. M.: Regression towards the mean, historically considered, *Statistical methods in medical research*, Vol. 6, No. 2, pp. 103–114 (1997).

[40] 高村大也：言語処理のための機械学習入門，言語処理のための機械学習入門 (2010).

© 2017 Information Processing Society of Japan