地理的損失を考慮したマルチクラス分類による地域推定

数原 良彦 †　　戸田 浩之 †　　鷲崎 誠司 †
† 日本電信電話株式会社 NTT サービスエボリューション研究所
†{suhara.yoshihiko, toda.hiroyuki, suzaki.seiji}@lab.ntt.co.jp

1　はじめに

スマートフォンやタブレット端末の普及に伴い，位置情報サービスの需要が高まっている．位置情報サービスは，ユーザの位置情報を利用してユーザの状況に応じたコンテンツを提供するサービスであり，i コンシェル[1] や Google Now [2] が挙げられる．これらのサービスでは，位置情報を元に検索結果や推薦結果を変えることにより，ユーザの状況に合わせた情報提示を行っている．

グルメサイトや施設情報サービスにおける店舗情報に関しては，サービス提供側に正確な位置情報が付与されたコンテンツが潤沢に存在しており，これを利用したサービス提供が可能である．また，ブログや Twitter をはじめとしたソーシャルメディアには，様々な背景を持ったユーザによる意見が記述されており，これらを分析し，活用することで位置情報サービスに施設情報とは異なる付加価値を付与することが可能だと考えられる．しかしながら，ブログや Twitter のようなソーシャルメディアにはほとんどのコンテンツに明示的な位置情報が付与されておらず，情報をそのまま位置に紐づけることができない．Cheng ら [1] によれば，Twitter において，緯度経度情報を含むジオタグが付与されたツイートは全体の 0.42%であると報告されており，多くのツイートが地理情報に紐づけられていない．そのため，ソーシャルメディアのコンテンツに対する位置情報推定ができれば，ユーザの生の声を地理情報に紐づける位置情報サービス提供が可能だと考えられる．

ソーシャルメディアのテキスト情報をもとに位置情報を推定する研究は多数行われている [1][2][3]．これらの方法では Twitter データを対象にし，あるユーザの発信した情報をひとつの大きなテキストをみなし，当該ユーザの居住地を正解ラベルとみなして，教師あり学習を用いたマルチクラス分類として解いている．Han ら [2] は特徴選択を行い，Information Gain に基づく特徴選択によって推定精度が向上することを示しており，テキスト情報のみを用いた位置推定に対する知見を与えている．評価指標としては，予測ラベルが正解ラベルに一致する正解率 (Accuracy) に基づく評価と，予測ラベルが正解ラベルの 10 マイル以内に存在するかという正解率 (Acc@161) に基づく評価方法が用いられている．

地理上の分布を分析するための目的で用いる地域推定においては，真の値域の近傍と判定する誤りが，遠くの地域と判定する誤りに比べて許容される場合が考えられる．このようなアプリケーションにおいては，既存研究で用いられた Acc@161 のように，地域推定タスクにおいては，距離を考慮した評価が現実的であり，完全一致以外を等しく誤りとする評価指標は想定されるアプリケーションに合っていないといえる．この場合，マルチクラス分類において，完全一致以外を等しく誤りとみなす学習アルゴリズムではこのような距離を考慮した評価指標を最大化する観点から適切ではない．

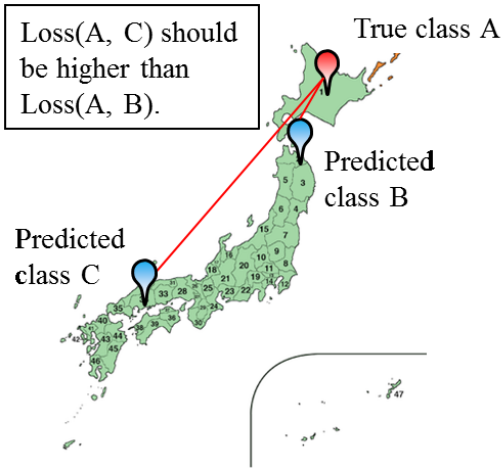

図 1: 本研究のモチベーション

既存手法ではマルチクラス分類器の学習フェーズにおいて，損失関数に地理的な観点を取り入れていないという問題が挙げられる．すなわち，日本国内における居住地推定の例では北海道に住むユーザの居住地を青森県と誤る損失と，広島と誤る損失が等しいという仮定に基づいている (図 1 参考)．これは不自然な仮定であり，アプリケーションの観点からも，予測ラベルが正解ラベルの近傍に存在するべきいう観点で評価することが自然である．本稿では，このような地理的な損失を考慮した学習が可能であるか，という Research Question の解決を目指す．

そこで本稿では，コスト考慮型識別学習において，コスト関数に地理的損失の観点を取り入れることにより，従来手法に比べて高精度な地域推定を目指す．我々の知る限り，テキストを用いた地域推定に地理的損失を考慮した識別学習アルゴリズムの開発を行った既存研究は存在しない．また，既存研究では MNB [1][2] の

[1] https://www.nttdocomo.co.jp/service/customize/iconcier/
[2] http://www.google.com/landing/now/

結果が報告されており，生成モデルと識別学習の性能比較が行われていない．そのため，本稿では生成モデルと識別学習の比較を通じて，本タスクにおける学習アルゴリズムの性能について得られた知見を述べる．

本稿の貢献は以下のとおりである:

- コンテンツベースの地域推定において，コスト考慮型マルチクラス分類を用いてラベル誤りに対して地理的損失を導入する方法を提案し，Multiclass PA に本アイディアを適用した Distance-Sensitive MPA を開発する．
- 既存手法で利用された手法と地理的損失を考慮しない手法との比較実験を通じて提案手法の有効性を検証する．

2 関連研究

ソーシャルメディアのコンテンツに対して位置情報を推定する研究は数多く研究されており，(1) 特徴抽出手法の提案 [1][3][4][5][6]，(2) 分類手法の提案 [7] の 2 つに分けられる．本稿における我々の提案は地理的損失を考慮したマルチクラス分類手法の実現であり，(2) に該当する．

Cheng ら [1] はツイート本文の情報のみを使い，地域 (市単位) に対して特徴的な単語を抽出することにより，居住地の推定を行っている．居住地の正解データは，ユーザのプロフィールの Location 欄を使用している．地理的な位置に応じてツイート本文内の単語の分布が異なることを利用し，ユーザのツイート情報のみを用いて位置推定を実現した．Chandra ら [4] もまた，Twitter ユーザの居住地の推定について述べている．Chandra らの手法では，本文情報の単語の頻度や，Reply と呼ばれるユーザ同士のやり取りの機能，また，やり取りを始めたユーザの位置情報を利用して推定を行っている．Hecht ら [5] も同様に，Twitter ユーザの居住地の推定を行っている．特徴的な単語を選択し，アメリカの都市と 4 カ国をそれぞれ分類している．橋本ら [6] は，ツイートから得られる環境情報を TF-IDF を用いて抽出する手法を提案している．ツイート本文に対して形態素解析器で単語分割を行い，TF-IDF を利用して，特徴的な単語を抽出した．また，TF-IDF の計算には単語の出現回数ではなく，エリアで発生したツイート内で単語を用いたユニークユーザ数を利用することにより，エリア内で大量のツイートをするユーザや特定の単語を繰り返しツイートするユーザの寄与を抑えて，特徴的な単語の抽出を行っている．西村ら [3] は橋本らの手法に基づいて単語の特徴選択を行い，Twitter ユーザの居住地推定を行い，TF-IDF に基づく特徴選択によって分類性能を向上することを示している．

安田ら [7] は，ブログのテキストからブログ作者の居住地を推定している．ブログ中に地名を含むブログ作者を対象にして，ブログ作者が居住しているかどうかを 2 値分類を行っている．分類器を複数用意し，すべての地名に対する 2 値分類結果の投票で判定を行っている．この手法は 57.6%の正解率で都道府県を推定している．安田らの研究では，地名を含むブログを対象にしており，地名の周囲の単語を素性としている．地域推定のための学習手法の提案という点で関連するが，安田らの手法は地域誤りの損失を考慮していない点で本研究と異なる．

3 マルチクラス分類による地域推定

本稿では，離散的な地域ラベル $y \in Y$ が与えられた事例集合 D が与えれた際に，未知の文書に対して地域ラベルを推定する問題を考える．Twitter においてユーザの居住地推定の例では，ユーザを事例，ユーザの居住地が地域ラベルと見なす．先述のとおり，ジオタグや IP アドレスに基づいて連続的な緯度経度情報を取得するコストは高く，離散的な地域ラベルが付与されている問題設定の方が現実的であると考えている．

本稿では，オンラインの識別学習手法である Passive-Aggressive (PA) [8] を位置推定に適したアルゴリズムに拡張する．本稿では (1) 大規模データに適用可能であり， (2) 追加学習容易なオンライン学習アルゴリズムであることの 2 点の理由からオンライン識別学習手法である PA を選択した．まず既存手法である Multiclass PA について述べたのちに，地理的損失を導入する提案手法について述べる．

3.1 Multiclass PA

マルチクラス分類 PA (Multiclass PA; MPA) では，試行 t における重みベクトル $\mathbf{w}_t$ を用いて，最大損失を与えるクラス

$$\hat{y}_t = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbf{w}_t^T \Phi(\mathbf{x}_t, y_t) - \mathbf{w}_t^T \Phi(\mathbf{x}_t, y) + \sqrt{\rho(y_t, y)}$$

を取得し，以下の損失を 0 とするような更新を行う:

$$\ell_t = \mathbf{w}_t^T \Phi(\mathbf{x}_t, y_t) - \mathbf{w}_t^T \Phi(\mathbf{x}_t, \hat{y}_t) + \sqrt{\rho(y_t, \hat{y}_t)}$$

ここで $\mathbf{x}_t$ は試行 t において選択された事例を表し，$\Phi(\mathbf{x}_t, \hat{y})$ はクラス y と事例 $\mathbf{x}_t$ に基づいた特徴ベクトルを返す素性関数を表す．また y_t は試行 t において選択された事例の真のラベルを表し，$\rho(y_t, \hat{y}_t)$ はラベル y_t をラベル $\hat{y}_t$ と予測する際のコストを表す．

3.2 Distance-Sensitive MPA

我々は，識別学習の損失関数に何かしらの形で地理的損失の情報を導入することができれば，コスト考慮型学習の枠組みで地理的損失を考慮した識別学習を実現することが可能であると考えた．そこで本稿では MPA における $\rho(y, \hat{y})$ に地理的損失を知識として導入する手法 Distance-Sensitive Multiclass PA (DSMPA) を提案する．地理的損失の設定方法として，本稿では以下の 2 つの方法を提案し，評価実験においてこれらの有効性を検証する．

(1) 地域隣接に基づく損失の設定
位置ラベルについて隣接する地域については損失を小さく設定し，それ以外の地域については大きな損失を設定する方法 ($\text{DSMPA}_{\text{Adj}}$).

(2) エリアに基づく損失の設定
真のラベルと同エリアに属する地域については損

Algorithm Distance-Sensitive MPA (DSMPA)

Input: $(\mathbf{x}_n, y_n) \in D$, T, C, $\rho(y_i, y_j)\,\forall i, j \in \{1, \ldots, |Y|\}$
Output: $\mathbf{w}^*$

1: **Initialize**: $\mathbf{w}_1 = \mathbf{0}$
2: **FOR** t in 1 to T
3: Uniformly choose class c.
4: Sample $(\mathbf{x}_t, y_t)$ where $y_t = c$.
5: $\hat{y}_t = \underset{y \in Y}{\operatorname{argmax}}\, \mathbf{w}_t^T \Phi(\mathbf{x}_t, y_t) - \mathbf{w}_t^T \Phi(\mathbf{x}_t, y) + \sqrt{\rho(y_t, y)}$
6: $\ell_t = \mathbf{w}_t^T \Phi(\mathbf{x}_t, y_t) - \mathbf{w}_t^T \Phi(\mathbf{x}_t, \hat{y}_t) + \sqrt{\rho(y_t, \hat{y}_t)}$
7: Calculate $\tau_t = \min\left\{C, \frac{\ell_t}{||\mathbf{x}_t||_2^2}\right\}$
8: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \tau y_t \mathbf{x}_t$
9: **ENDFOR**
10: $\mathbf{w}^* = \frac{1}{T}\sum_{i=1}^{T} \mathbf{w}_i$
11: **RETURN** $\mathbf{w}^*$

図 2: Distance-Sensitive MPA

表 1: 本稿で用いる同一エリアの都道府県

地方	都道府県名
北海道	北海道
東北	青森県 岩手県 宮城県 秋田県 山形県 福島県
関東	茨城県 栃木県 群馬県 埼玉県 千葉県 東京都 神奈川県
中部	新潟県 富山県 石川県 福井県 山梨県 長野県
東海	岐阜県 静岡県 愛知県 三重県
近畿	滋賀県 京都府 大阪府 兵庫県 奈良県 和歌山県
中国	鳥取県 島根県 岡山県 広島県 山口県
四国	徳島県 香川県 愛媛県 高知県
九州	福岡県 佐賀県 長崎県 熊本県 大分県 宮崎県 鹿児島県 沖縄県

失を小さく設定し，それ以外の地域については大きな損失を設定する方法 ($\text{DSMPA}_{\text{Area}}$).

図 2 に DSMPA のアルゴリズムを示す．各試行においてデータセット D に含まれるクラス集合 Y からランダムにクラスを選択し (ステップ 3)，選択されたクラスから事例 $\mathbf{x}_t$ を 1 つランダムに選択する (ステップ 4)．これは各クラスの事例偏りの影響を排除するために各クラスから等しくサンプルするというヒューリスティクスである．選択された事例の正解ラベル y_t と，あらかじめ与えられたコスト関数 $\sqrt{\rho(y_t, y)}$ に基づいて損失を計算し，最大損失を与えるクラス $\hat{y}$ を選択する (ステップ 5)．クラス $\hat{y}$ によって与えられる損失に基づいて更新量 τ_t を計算する (ステップ 7)．この際，事前に設定された C パラメータで更新量を抑える PA-I の戦略を用いる．以上の処理を事前に設定された試行 T 回行い，イテレーションの最後に重みベクトルの平均化 [9] を行う (ステップ 11)．

4 評価

提案手法の有効性を検証するため，評価実験に Twitter データを利用し，ユーザの居住地推定タスクの評価を行った．地域ラベルとして日本における 47 都道府県を対象とした．すなわち本実験は 47 クラスのマルチクラス分類である．2012 年 1 月 1 日から 6 月 30 日までの Twitter Stream API を利用して取得した Public timeline 全体の約 10%のツイートデータを対象とした．期間内のツイートデータのうち，Twitter の言語設定が日本語表示であり，かつ Location フィールドに 47 都道府県名の文字列を含むユーザを対象とし，6ヵ月間で 50 ツイート以上発信している計 114,522 ユーザを対象とした．この際，Location フィールドに記述された都道府県名を当該ユーザの居住地ラベルとして利用した．各ユーザの期間内のツイートを当該ユーザに対応する文書とみなし，これらのツイートについて形態素解析器 JTAG [10] を利用して形態素解析を行った．95%以上のユーザに使用されている単語，使用されているユーザが 5 人未満の単語はストップワードとして除去し，残りの単語の bag-of-words を特徴として用いた．

4.1 実験条件

比較アルゴリズムとして既存手法 [1][2] で用いられた Multinomial Naive Bayes (MNB)，MPA，$\text{DSMPA}_{\text{Adj}}$，$\text{DSMPA}_{\text{Area}}$ の 4 つの比較を行った．提案手法における $\rho(y, \hat{y})$ 設定の妥当性を検証するため，MPA については図 2 において $\rho(y, \hat{y}) = 0\,\forall y, \hat{y} \in Y$ としたものを用いた．

本データセットにおける DSMPA におけるコスト設定について述べる．地域隣接に基づく損失を用いる $\text{DSMPA}_{\text{Adj}}$ には，47 都道府県の隣接に基づく情報を利用した．隣接は県境を挟んで隣り合う場合に都道府県が隣接していると判断し，県境が海上に存在する場合も含めた．真のクラス y に対して隣接県クラス $\hat{y}_{adj}$ に誤るコスト $\rho(y, \hat{y}_{adj}) = 0$ とし，隣接県以外のクラス $\hat{y}_{\neg adj}$ に誤るコストは $\rho(y, \hat{y}_{\neg adj}) = 2$ と設定した．

エリアに基づく損失を用いる $\text{DSMPA}_{\text{Area}}$ には北海道・東北・関東・中部・近畿・中国・四国・九州に分ける八地方区分を用いた．八地方区分を表 1 に示す．エリアに基づく損失においては，真のクラス y に対して同一エリアであるクラス $\hat{y}_{area}$ に誤るコスト $\rho(y, \hat{y}_{area}) = 0$ とし，同一エリア以外のクラス $\hat{y}_{\neg area}$ に誤るコストは $\rho(y, \hat{y}_{\neg area}) = 2$ と設定した．

データセットに含まれる Tweet をユーザ単位で 5 分割し，3 ブロックを訓練データ、1 ブロックを検証データ，1 ブロックをテストデータとして利用する評価を 5 回行う 5 分割交差検定で評価を行った．MPA と DSMPA における C パラメータ ($C \in \{10^{-5}, 10^{-4}, 10^{-3}, \ldots, 1.0, 10.0\}$) およびイテレーション数 T ($T \in \{10^5, 5 \cdot 10^5, 10^6\}$) は検証データにおいて評価指標が最大となる値を選択した．MNB におけるスムージングには加算スムージングを利用し，単語頻度に加算する α パラメータ ($\alpha \in \{10^{-5}, 10^{-4}, 10^{-3}, \ldots, 1.0\}$) は，MPA や DSMPA と同様に検証データにおいて評価指標が最大となる値を選択した．

評価指標には正解率 (Acc)，隣接県の誤りを正解とみなした正解率 (AdjAcc)，同一地方を正解とみなす正解率 (AreaAcc)，そして一定距離以内の誤りを正解とみなす正解率である Acc@d を用いた．Acc@d には既存研究で用いられた $d = 161$ の他に，$d = 50, 80.5$ を用いた．なお，本実験では都道府県をクラスとしているため，クラス間の距離が定義されていない．そのため，本実験では都道府県間の距離として各都道府県の県庁所在地間の距離を用いて d の判定に用いた．またクラス誤りについて平均距離誤り (Dist Loss) も用

表 2: 実験結果

Method	Acc	AdjAcc	AreaAcc	Acc@50	Acc@80.5	Acc@161	Dist Loss
MNB	0.234	0.350	0.468	0.408	0.433	0.507	−294.7
MPA	0.242	0.351	0.480	0.404	0.435	0.523	−279.1
$\text{DSMPA}_{\text{Adj}}$	**0.256**	**0.377**	**0.517**	**0.441**	**0.476**	**0.555**	**−250.6**
$\text{DSMPA}_{\text{Area}}$	0.250	0.370	0.506	0.432	0.465	0.542	−266.38

いた．これは県庁所在地間の距離を距離誤りとして利用し，予測あたりの平均値である．値が大きい方が性能が高いという評価指標に対する説明の一貫性を保つため，距離誤りには負の符号を付与する．

4.2 結果と考察

実験結果を表 2 に示す．$\text{DSMPA}_{\text{Adj}}$ が全ての評価指標において最大の値を示した．これにより，提案手法の有効性を確認するとともに位置誤りを考慮した学習により，高精度な位置推定が可能であることを確認した．

$\text{DSMPA}_{\text{Adj}}$ と $\text{DSMPA}_{\text{Area}}$ が地域誤りを考慮した評価指標において高い値を示した．これにより，学習における損失に地域の隣接情報を導入する方法，同一エリアに対して等しい損失を与える方法いずれの場合においても位置誤りを考慮した学習を実現しているといえる．

実験結果より，位置誤りを考慮した評価指標だけでなく，正解率についても地理誤りを考慮した $\text{DSMPA}_{\text{Adj}}$ と $\text{DSMPA}_{\text{Area}}$ が MPA に比べて高い値を示した．これより，Twitter ユーザの居住地推定というタスクにおいては，クラスに対する地理的情報が分類精度向上に寄与するといえる．これは，特徴空間上における各クラスの事例分布同士が，地理的関係を持っているということが推測される．

MNB は Acc@161 で 0.507 の値を示したが，これは Cheng ら [1] の Acc@161=51%という結果と大体一致する．データセットと実験条件が異なるため直接の比較ができないものの，この結果より，米国における市ラベルの地域推定と，日本における都道府県推定において同じ学習アルゴリズムで同程度の精度を示すことを確認した．

また実験結果より，MPA が MNB に比べて Acc@50 を除く評価指標において僅かに高い値を示した．これにより，本タスクにおいては MPA が MNB に比べて優れていることが示され，本タスクにおいては識別手法が生成手法に比べて有効に働くことが示唆された．

5 おわりに

本稿では，テキストのみを手がかりにした位置推定タスクにおいて，位置誤りを考慮した評価指標を最大化する方法を検討し，クラスの誤り損失に対して地理的知識を導入するオンライン学習アルゴリズムである Distance-Sensitive Multiclass PA を提案した．Twitter データセットのユーザ居住地推定タスクにおける評価実験を通じて，既存手法で用いられた Multinomial Naive Bayes，地理的知識を利用しない Multiclass PA に比べて，正解率，位置誤りを考慮した正解率共に高い精度で分類可能なことを示した．

参考文献

[1] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc CIKM '10*, pp. 759–768, 2010.

[2] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proc. COLING 2012*, pp. 1045–1062, 2012.

[3] 西村駿人, 数原良彦, 鷲崎誠司. 地域特徴語選択を用いたマルチクラス分類による twitter ユーザの居住地推定. 第 4 回集合知シンポジウム NLC2012-37, pp. 23–27, 2012.

[4] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions–a content based approach. In *Proc. PASSAT/SocialCom '11*, pp. 838–843. IEEE, 2011.

[5] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proc. CHI '11*, pp. 237–246, 2011.

[6] 橋本康弘, 岡瑞起. 都市におけるジオタグ付きツイートの統計. 人工知能学会誌, Vol. 27, No. 4, pp. 424–431, 2012.

[7] 安田宜仁, 平尾努, 鈴木潤, 磯崎秀樹. ブログ作者の居住域の推定. 第 12 回言語処理学会年次大会, pp. 512–515, 2006.

[8] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithm. *Mach. Learn.*, Vol. 7, pp. 551–585, 2006.

[9] Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proc. EMNLP '02*, pp. 1–8, 2002.

[10] Takeshi Fuchi and Shinichiro Takagi. Japanese morphological analyzer using word co-occurrence: JTAG. In *Proc. COLING '98*, pp. 409–413, 1998.