ISS-P-033

ツイートを対象としたノンパラメトリック検定でのクラスタ細分化による属性推定

吉高　太志†　　田村　慶一†　　森　康真†
† 広島市立大学大学院情報科学研究科

1. はじめに

SNS はユーザの意思決定に関与すると考えられ，SNS で広告を見かけることが一般的になった．このような広告は，性別についてターゲットを絞ることがある．よって，SNS における性別推定はマーケティングに活用される．[1]は，比較実験で使用する n-gram や素性である文字・単語・品詞についての先行研究である．本研究では，ツイートを対象とした Twitter ユーザの性別推定について，ノンパラメトリック検定によりクラスタの細分化した性別推定を提案する．

2. 従来手法

従来手法の流れを次に示す．

1. Twitter のアカウントを人の目で性別推定し，アカウントごとにツイートを取り出す．
2. 全てのツイートを単語・品詞・文字に分割及び変換する．
3. n-gram を用いて，単語を特徴数値ベクトルに変換する．
4. 特徴数値ベクトルを，男性クラスタを0，女性クラスタを1として二種類のラベル付けを行う．
5. 2分類問題としてランダムフォレスト（以下 RF と記す）や SVM で性別推定を行う．

次に，この手法による精度を次の表1に示す．

表1．従来手法による性別推定の精度

	uni-gram					
	RF			SVM		
	女性	男性	合計	女性	男性	合計
単語	1	1	1	1	1	1
品詞	0.607	0.759	0.683	0.759	0.645	0.702
文字	0.721	0.556	0.639	0.721	0.696	0.708
	tri-gram					
	RF			SVM		
	女性	男性	合計	女性	男性	合計
単語	0.797	0.506	0.651	0.632	0.746	0.689
品詞	0.594	0.835	0.715	0.715	0.696	0.689
文字	0.746	0.721	0.734	0.734	0.696	0.721

3. 提案手法

提案手法では従来手法で示した実験の流れの4を削除し，次に示す 4´を加えた．

4´．男性・女性クラスタに関して，更にクラスタを分割する．

3.1 クラスタの細分化

クラスタの細分化は以下のようにして行った．

1. 閾値 A を設定する．
2. 2つの特徴数値ベクトルに関して，コルモゴロフースミルノフ検定を行い，二つの経験分布の類似度 S を参考値とする．
3. 閾値Aの範囲内である類似度Sについて，値の小さいものから順に，対応する2つの特徴数値ベクトルのラベルを統一する．

テストデータに関して，RF や SVM で性別推定した際，推定されたクラスがもともと男性であれば男性に予想されたと考える．

3.2 コルモゴロフ―スミルノフ検定

コルモゴロフースミルノフ検定とノンパラメトリック検定の一種で，正規分布していなくても使える，経験分布の形状に対する検定である．

本研究ではクラスタを細分化する際，コルモゴロフースミルノフ検定の示す面積を分布の類似する度合いと見立てて用いた．

4. データセット

男性・女性アカウントを人手で79アカウント収集した．

性別推定基準は，以下の項目を評価し，総合的に推定した．

・Twitter プロフィールや Facebook などのリンク先にある，実名と思われる登録名や性別登録欄の性別
・Twitter の登録名・自己紹介欄・ツイート内容
・ツイート診断アプリでの性別推定結果や，その診断結果に対するユーザのコメントや，診断結果に対する知人と思われるフォロワーのコメント

5. むすび

今後の展望を次に記す

・提案手法による，精度の向上に資する研究をする．
・閾値の値を変えてみて，最適なパラメータを探す．
・クラスタの分割条件を変化させる．
・他の素性でも提案手法を試す．
・男女推定に重要な単語等の確認をする．

参考文献

[1]財津亘 ほか，ランダムフォレストによる著者の性別推定 －犯罪者プロファイリング実現に向けた検討－ ，情報知 識学会誌Vol. 27 No. 3 , pp. 261-274 , 2017