D-012

Tweet の分類による観光情報の取得
Collecting Tourism Data by Classifying Tweets

関谷 虎汰郎 [1)]　鈴木 優[1)]
Kotaro Sekiya　Yu Suzuki

1　はじめに

観光は日本の経済において重要な成長分野というだけでなく，地域経済活性化の推進力と位置付けられている．現在，観光の実態として観光客は大都市に多く集まり，地方都市への訪問は限られている．しかし，近年 ICT の利用などによる集客の期待が高まっている．また，SNS 利用者の増加により，投稿された情報を分析する研究が行われている．その中の一つに観光に着目したものがある．SNS 上の豊富な情報を利用することにより，多くの観光情報を収集できることが期待される．しかし，投稿される情報には観光以外の情報も存在する．そのため，観光情報を含んでいるかどうかの分類が必要である．そこで，機械学習による分類を考える．多くの教師あり学習において，適切に機能するには多くのラベル付けされたデータが必要である．十分なデータを用意するには費用や時間といったコストが多く必要である．そこで，より少ない情報量で精度を期待できる能動学習を用いた分類とクラウドソーシングを利用したラベル付けによりコストを抑えることができると考えた．また，クラウドソーシングの活用により，システムの精度の向上を目指す．今回，能動学習の効果の確認とクラウドソーシングとの対比のため，第一著者によるラベル付けにより実験を行った．

2　関連研究

能動学習とは，機械学習の枠組みの一つである．学習者が自ら学習すべきデータであるかを判別することでラベル付けデータの取得コストを抑えながら精度を出すことのできる学習方法である．能動学習はさまざまな手法が考案されており，Settles[1] はその手法を分類し，詳細を記載している．また，クラウドソーシングと機械学習を用いる研究も行われている．Rayker ら [2] は，CAD の情報解析のタスクにおいて，ワーカのラベル付けの正確さを考慮に入れた学習方法を提案している．一方本研究では，Tweet の分類による岐阜の観光情報の取得というタスクにおいて，ワーカ間のラベル付けの一致度を用いることで，有用性の高い情報の収集が可能だと考えた．そこで，今回は事前実験としてラベル付けを第一著者が行い，その精度を確認した．

3　分類手法

図 1 にシステムの全体図を示す．

3.1　前処理

(1)TwitterAPI を用いて Tweet を収集する．
(2)収集した Tweet を第一著者により人手で観光情報を含むかどうかを分類し，ラベル付けを行う．
(3)収集した Tweet から記号と数字，Tweet 内でよく見られ，観光情報と関係のないと考えられる “RT”，“お気に入り”，“Tweet”，“Web”，“まとめ” を取り除く．その後，MeCab を用いて形態素解析を行う．辞書は IPA 辞書と NEologd 辞書とする．特徴語として名詞を抽出し，品詞細分類で非自立と判定された語は取り除き，教師データとする．
(4)新たに収集した Tweet に (3) と同様の処理をし，入力データとする．

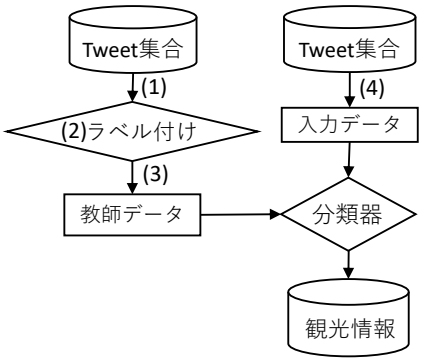

図 1　Tweet の分類の流れ

3.2　分類器

以下五つの手法を実装し，入力データが観光情報を含むかどうかの分類を行う．このとき手法 4 以外では，入力データを TF-IDF を用いてベクトル化をする．

- 手法 1：Random Forest
- 手法 2：SVM(Support Vector Machine)
- 手法 3：ロジスティック回帰
- 手法 4：関連フィードバック
- 手法 5：Uncertainty Sampling

3.2.1　手法 1〜3: 教師あり学習

手法 1〜3 は十分なラベル付きデータの必要な教師あり学習である．能動学習との精度比較，コストの比較に用いる．

3.2.2　手法 4: 関連フィードバック

ラベルごとに Tweet 内の名詞を集計し，名詞群を作成する．この時，両ラベルに含まれる名詞が存在する場合，その名詞を含む Tweet の件数が多かったラベルの名詞として集計する．各入力データ内の全ての名詞のうち，名詞群の名詞と一致する名詞の占める割合をそれぞれ求め，分離境界を決定することで分類することを関連フィードバックとする．

3.2.3　手法 5: Uncertainty Sampling

Uncertainty Sampling は能動学習の一つである．データに対し常に正しいラベルを付ける存在を準備し，この存在をオラクルと呼ぶ．少量のラベル付けされたデータセットから学習を開始する．まだラベル付けされていないデータを学習者がラベル付けを行ったとき，ラベル付けがあいまいなデータを学習に有用であるデータとする．有用なデータのラベルをオラクルに問い合わせることで取得し，新たにデータセットに追加する．有用なデータの追加を繰り返したデータセットを元に入力データを分類することにより，精度の向上を目指した手法で

1)　岐阜大学 工学部 電気電子・情報工学科

ある．今回は，学習者のラベル付けにロジスティック回帰を用いた．オラクルを第一著者とし，ラベルなしデータを分類したときの予測確率の差が少ないものを問い合わせ，データセットに追加する．その後，追加されたデータセットを用いて新たにロジスティック回帰を学習させ，入力データの分類に用いる．

4 評価実験

この実験では，能動学習により，ラベル付けコストを抑えながら実用的な精度が出ることを確認する．

4.1 手順

検索ワードを “岐阜城” など，岐阜の観光に関連のあるものとし，2020 年の 1 月上旬から 5 月上旬までの Tweet を収集する．人手による分類により，観光情報を含む Tweet を 893 件，観光情報を含まない Tweet を 1,000 件の計 1,893 件を用意し，7 割を訓練データ，3 割をテストデータとする．手法 1〜3 は scikit-learn を用いて実装する．また，5 分割交差点検証を行い，パラメーターを決定する．手法 4 では，訓練データを用いて名詞群を作成し，テストデータの割合を求める．手法 5 は alipy を用いて実装する．訓練データの 7 割のうちさらに 3 割を初期データセット，7 割をラベルなしデータとして使用する．初期データ，ラベルなしデータ，テストデータの組み合わせを変更して 5 回学習し，精度の平均を出す．1 度の問合せで追加されるデータを一つとする．

4.2 実験結果

表 1〜4，図 2，3 において，正解データを正解，観光情報を含む Tweet を含む，観光情報を含まない Tweet を含まないとそれぞれ表記する．手法 1 では深さの最大値を 38 とし，accuracy は 0.867，手法 2 ではカーネルを rbf，$C = 5$ とし，accuracy は 0.892，手法 3 では，$C = 6$ とし，accuracy は 0.908 となった．また，それぞれの混同行列を表 1，表 2，表 3 に示す．手法 4 の各テストデータごとの割合をプロットしたグラフを図 2 に示す．ここで，x 軸は観光情報を含まない Tweet から集計した名詞群の名詞と一致した名詞が占める割合，y 軸は観光情報を含む Tweet から集計した名詞群の名詞と一致した名詞が占める割合である．この図より，ラベル間のプロットに明確な差異はなく，今回のタスクにおいて，手法 4 は十分な分類精度が期待できないことが考えられる．手法 5 では，$C = 11$，問合せ回数 609 回で accuracy は 0.912 となった．この時の混同行列を表 4 に，問合せ回数と accuracy の推移を図 3 に示す．また，問合せ回数が 477 回で手法 3 と同程度の精度が見込めることから，少ないラベル付きデータから実用的な精度が得られると考えられる．

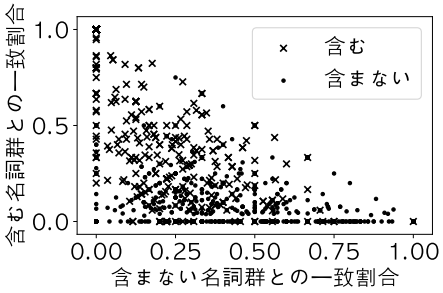

図 2 関連フィードバック

表 1 Random Forest の混同行列

		分類結果	
		含む	含まない
正解	含む	289	11
	含まない	64	204

表 2 SVM の混同行列

		分類結果	
		含む	含まない
正解	含む	288	12
	含まない	49	219

5 おわりに

本研究では，Tweet の分類による観光情報の取得というタスクにおいて，能動学習の精度について検証した．実験では，能動学習はラベル付けのコストを抑えながら，実用的な精度が出ることが分かった．これにより，能動学習の有用性を確認できた．今回は人手でのラベル付けを第一著者のみで行ったため，ラベル付けの妥当性，クラウドソーシングの有用性を示すことが出来なかった．今後はクラウドソーシングを組み合わせることにより，コストの軽減，精度の向上を目指す．

謝辞

本研究の一部は JSPS 科研費 18H03342，19H04221，19H04218，および大川情報通信基金の助成を受けたものです．

参考文献

[1] Burr Settles．Active Learning Literature Survey．Computer Sciences Technical Report 1648．University of Wisconsin–Madison．Updated on: January 26, 2010.

[2] Vikas C. Raykar，et al.，Learning From Crowds．Journal of Machine Learning Research, Vol. 11, pp. 1297–1322 (2010).

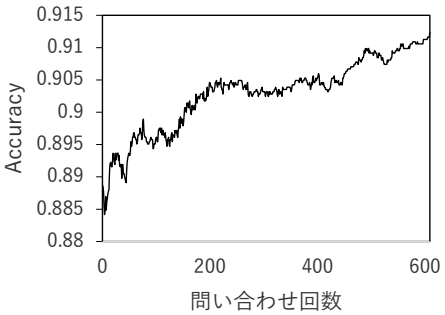

図 3 Uncertain Sampling

表 3 ロジスティック回帰の混同行列

		分類結果	
		含む	含まない
正解	含む	283	17
	含まない	35	233

表 4 Uncertain Sampling の混同行列

		分類結果	
		含む	含まない
正解	含む	288.4	13.6
	含まない	36.2	229.8