3ZD-06

SNS 情報収集支援における深層学習の応用

山崎　拓己†　岡本　東†　堀川　三好†
岩手県立大学ソフトウェア情報学部†

1. はじめに

SNS (Social Networking Service) 利用者の増加に伴って, 投稿を活用した広告・宣伝や市場分析に関する研究が数多くなされている. しかしながら, 投稿収集者の目的に合った投稿を的確かつ効率的に収集することは難しい. 本研究の目的は投稿に対する収集者の評価情報を学習した投稿分類システムを提案し, 投稿情報活用の支援を行うことである. 本稿では滝沢市観光ポータルサイト（以下観光ポータル）で収集している Twitter の投稿を対象に, 順伝播型ニューラルネットワーク（以下 FNN）を用いた投稿分類を検証した結果について報告を行う.

2. 関連研究

機械学習を用いて投稿の分類を行う研究は多くなされている[1]. 近年では, テキスト分類において, 深層学習を用いた分類手法が深層学習以外を用いた分類手法よりも精度が高いことに注目されている[2].

3. 分類機能の要件

本稿で対象としている観光ポータルでは約 180 個存在する観光資源の PR を目的に Twitter から投稿を収集している. 現在, ポータルでは観光資源名を Twitter API で検索した結果を保存しており, すべての投稿を観光資源の PR に活用することができていない. したがって, 分類システムは集められた投稿を PR に活用できるものとそれ以外とに分類する必要がある. また, 投稿文は一般的な文語文と比較したときに未知語や顔文字を多く含むため, 分類システムそれらに対してロバストでなければならない.

4. 分類手法

4. 1. 訓練データの作成

訓練データは収集者が Tweet を観光資源の PR に活用できるものとそうでないものに手動で分類する.

4. 2. 訓練データの前処理

機械学習を用いて投稿文を分類するためには文の特徴を表現するベクトルを得る必要がある. 文の前処理として BOW (Bag-of-Words) が挙げられるが本提案では BOC (Bag-of-Characters) を用いる. BOC は形態素解析を必要とせず未知語等を多く含む投稿文において BOW よりロバストであり, 投稿の前処理として適当である.

4. 3. FNN の学習

FNN の学習では, はじめに自己符号化器を用いた訓練データの教師無し学習を行う. 次に, 訓練データの教師あり学習を行う. 教師あり学習は収集者が訓練データに付与した評価と FNN の出力との平均二乗誤差が最小になるように学習を行う. また過学習を抑えるドロップアウトを用いる.

5. 実験概要

5. 1. 実験条件

実験環境を表 1 に, FNN のパラメータを表 2 に示す. データセット（以下 DS）は観光ポータルで収集されている岩手山に関する投稿と岩手県立大学に関する投稿を対象とする. 岩手山 DS は訓練データが真偽ともに 100 件であり, 偽のデータの内 50 件が岩手山以外の投稿である. テストデータは真偽ともに 600 件で, 偽のデータの内 350 件が岩手山以外の投稿である. 岩手県立大学 DS は訓練データ, テストデータともに真偽それぞれ 100 件となっている. また, いずれの実験でも自己符号化器を用いた事前学習を 10 回, 教師あり学習を 100 回行う.

表 1　実験環境

OS	Windows10
プロセッサ	Intel core i5-4590
メモリ	8GB
GPU	GeForce GTX 750 Ti
IDE	Visual Studio Community 2015
ライブラリ	Chainer, Scikit-learn

表 2　FNN のパラメータ

層数	1
ユニット数	2048
ドロップアウト率	0.5
事前学習回数	10
教師あり学習回数	100
活性化関数	正規化線形関数

Deep learning application for gathering social media information.
†Takumi Yamazaki, Azuma Okamoto, Mitsuyoshi Horikawa , Iwate Prefectural University, Faculty of Software and Information Science

5. 2. 比較手法

FNN の比較対象として Support Vevtor Machine（SVM）, Naïve Bayes（NB）, Logistic Regression（LR）を選択する. この三つのパラメータは機械学習ライブラリ scikit-learn の各手法に対応するモジュールのデフォルトの値を用いた.

5. 3. 学習方法の比較

岩手山 DS を用いて確率的勾配降下法(SGD), Momentum, AdaGrad, Adam の 4 つの学習方法を比較する. 学習方法以外の FNN のパラメータを等しく設定し, 教師あり学習中の最も高い F 値を結果として採用する. また, この 4 つの手法のパラメータは深層学習ライブラリ Chainer で用意されているデフォルトの値を用いる.

5. 4. 前処理方法の比較

岩手山 DS を用いて BOC と BOW の比較を行う. FNN のパラメータを等しく設定し, 教師あり学習中の最も高い F 値を結果として採用する. SVM, NB, LR との比較も行う.

5. 5. 岩手県立大学 DS の分類

岩手山 DS を対象に行った実験結果をもとに選択した学習方法と前処理方法を用いて岩手県立大学 DS の分類を行う. SVM, NB, LR との比較も行う.

表 3　BOC 分類結果

	適合率	再現率	F 値
FNN	0.8336	0.8350	0.8343
SVM	0.7916	0.7533	0.7720
NB	0.7603	0.7400	0.7500
LR	0.7778	0.7467	0.7619

表 4　BOW 分類結果

	適合率	再現率	F 値
FNN	0.7766	0.8867	0.8280
SVM	0.7980	0.6917	0.7411
NB	0.7869	0.7817	0.7842
LR	0.7762	0.6650	0.7163

表 5　岩手県立大学 DS の分類結果

	適合率	再現率	F 値
FNN	0.8246	0.9400	0.8785
SVM	0.7257	0.8200	0.7700
NB	0.6614	0.8400	0.7401
LR	0.6923	0.7200	0.7059

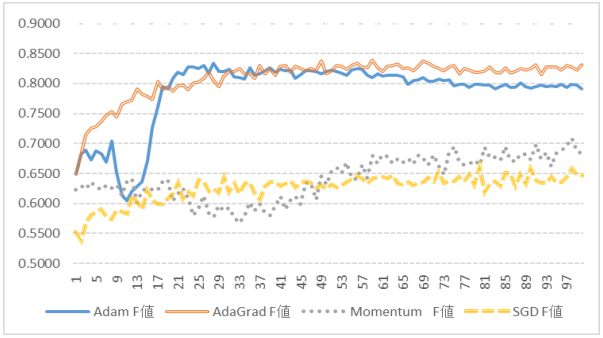

図 1　学習手法の比較（F 値）

6. 実験結果と考察

学習方法の比較結果を図 1 に前処理毎の分類結果それぞれ表 3 と表 4 に示す. 図 1 から Adam を用いた学習は収束が早いことがわかる. 今回の実験では AdaGrad の精度が高かったが 100 回の学習の内に収束しなかったため Adam を採用した. 前処理の比較では FNN において BOC を用いたときの適合率と F 値が BOW を用いたときより高いことがわかった. また, FNN を用いた分類が SVM 等よりも精度が高いこともわかった. 二つの実験結果を踏まえ, 学習方法に Adam, 前処理に BOC を採用して岩手県立大学 DS を分類した結果を表 5 に示す. 岩手山 DS の分類結果と同様の結果が得られた. これらの実験から前処理に BOC を用いた FNN の分類結果は他の分類結果と比較して適合率が高くなることがわかる. これは BOC が BOW と比べ, 分類に必要な特徴をよく表現しているからだと考えられる.

7. おわりに

本稿では SNS の投稿収集を支援するための深層学習を用いた投稿分類システムを観光ポータルに適用するための分類機能検証を行った. 結果から BOC の有用性と FNN の分類精度が高いことがわかった. 今後の課題として, 実際に投稿収集者が作ったテスト用データでの精度検証と Twitter 以外の投稿に対する検証が必要である.

参考文献

1) 山本修平, 佐藤哲司：Twitter からの実生活情報の抽出法の提案, データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2012）, (2012)
2) Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao: Recurrent Convolutional Neural Networks for Text Classification, 2015, Association for the Advancement of Artificial Intelligence 2015