

豊田工業大学 卒業研究報告書要旨

学部・学科	工学部 先端工学基礎学科		
学 籍 番 号	12056	氏 名	外山 洋太
題 名	文書・文間及びカテゴリ間の関係を考慮したレーティング予測		

1 序論

企業において、商品の評判分類は評判分析の中でも重要な問題である。従来の商品レビューを用いた複数カテゴリにおける評判分類手法では文間やカテゴリ間の関係を十分に考慮できていない。本研究は、複数カテゴリにおける評判分類について、文書及び文間の関係とカテゴリ間の関係を同時に考慮した分類を実現することを目的とする。

2 関連研究

2.1 隠れ状態を用いたホテルレビューのレーティング予測

藤谷ら [1] は複数のカテゴリにおける評判分類問題に対して、Multi-Instance Multi-Label learning for Relation Extraction (MIML-RE) [6] モデルを用いた手法を提案している。その手法では、文毎のレーティングからレビュー全体のレーティングを予測する際のカテゴリ間の繋がりを手動で変化させカテゴリ間の関係性を考慮している。

この手法では、文同士の位置関係を考慮しておらず、カテゴリ間については考慮しているものの深い関係性を捉えることができていない。

2.2 パラグラフベクトル

パラグラフベクトルは、文や文章といった大きな単位の言語表現の分散表現を学習する手法である。その手法の一つ、Distributed Memory model of Paragraph Vectors (PV-DM) によって得られたパラグラフベクトルは評判分類問題において高い正答率を示すことが示されている。しかし、文書全体にパラグラフベクトルを用いる場合、文同士の位置関係が分類時に考慮できない。

2.3 ニューラルネットワークを用いた評判分析

ニューラルネットワークを用いた評判分析の手法が、Nal ら [3]、Rie ら [4]、Duyu ら [5] 等に

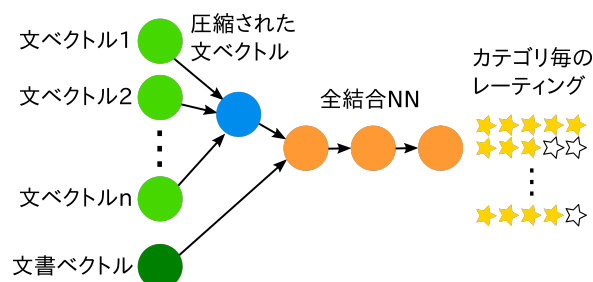


図1 全結合ニューラルネットワークによる分類器

よって提案されている。これらの方法に共通するのは、単語の意味表現から畳み込みニューラルネットワークと全結合ニューラルネットワークを用いて分類を行うことである。

これらの手法は1つのカテゴリにおける多値または二値分類を対象としている。よって、多カテゴリの評判分類問題において、これらの手法をカテゴリ毎に適用しただけではカテゴリ間の関係を考慮することができない。

3 提案手法

提案手法では、PV-DM によってレビュー内の文書全体及び各文の分散表現を生成し、それらをニューラルネットワークの入力として分類を行う。文毎の分散表現を用いることで文同士の位置関係を考慮し、ニューラルネットワークによる分類器を用いることで文間及びカテゴリ間の深い関係性を捉えることを目指す。提案手法の入力はレビューである文書と正解レーティングの組の集合、出力は各文書について予測されたカテゴリ毎のクラスである。

分類器の入力となる各レビュー内の文ベクトルは、各レビュー毎に重み付け平均によって圧縮する。この過程により、各レビューで疎らだった文の数を統一する。

分類器は全結合ニューラルネットワークによって構成される。図1に各層の結合の様子を示す。入力層はレビュー毎の文書ベクトルと圧縮された文ベクトルの結合ベクトルである。ま

た、出力層はカテゴリの数とラベルの数の積だけのユニットを持つ。

4 実験

4.1 実験設定

実験では、各手法の正答率を測定した。比較手法として、(1) Quoc ら [2] による PV-DM、及び、提案手法における分類器の入力を変えた2つの手法、(2) ASV (Averaged Sentence Vector) と (3) Weighted ASV を用いた。ASV ではレビュー内で平均した文ベクトルを、Weighted ASV ではレビュー内で重み付け平均によって圧縮された文ベクトルを分類器の入力として用いた。

データセットとしては、先行研究 [1] と同様に、ホテル予約サイト楽天トラベルにおけるレビュー 337,266 件からレビューの番号順に訓練データ 300,000 件、開発データ 10,000 件、評価データ 10,000 件を用いた。

4.2 結果と考察

まず、提案手法と3つの比較手法、従来手法 [1] を正答率で比較したものを表1に示す。提案手法が従来手法 [1] の正答率を 0.0189 上回っていることから、提案手法が従来手法 [1] より正答率において優れていることが分かった。また、Weighted ASV の正答率が ASV の正答率を 0.0018 上回っていることから、文の位置関係の考慮がレーティング予測に有効であることが分かった。さらに、提案手法が PV-DM や Weighted ASV に比べ高い正答率を示していることから、文書ベクトルと文ベクトルを同時に特徴量として用いることがレーティング予測に有効であることが分かった。これは文書ベクトルと文ベクトルがいくらか異なる特徴を学習していることを示す。

5 結論

本研究では、多カテゴリにおける評判分類問題について、レビュー全体の文書ベクトルに加え重み付け平均された文ベクトルを用いた手法を提案した。

実験では、従来手法 [1] に比べ提案手法が 0.0189 高い正答率を示した。また、提案手法

が比較手法よりも高い正答率を示したことから、レビュー内の文の並びが評判分類に重要であること、及び、文書ベクトルと文ベクトルがレビューのいくらか異なる特徴を捉えていることが分かった。

今後の課題は、文書や文の分散表現を生成する過程をニューラルネットワークによる分類器に統合することである。これによって、学習方法の柔軟性を高めると共にさらなる正答率の向上を目指す。

参考文献

- [1] 藤谷宣典ら, 隠れ状態を用いたホテルレビューのレーティング予測. 言語処理学会第21回年次大会, 2015.
- [2] Quoc Le, and Tomas Mikolov, Distributed Representations of Sentences and Documents. ICML 2014, 2014.
- [3] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom, A Convolutional Neural Network for Modelling Sentences. ACL 2014, 2014.
- [4] Rie Johnson, and Tong Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. NAACL 2015, 2015.
- [5] Duyu Tang, Bing Qin, and Ting Liu, Learning Semantic Representation of Users and Products for Document Level Sentiment Classification. ACL 2015, 2015.
- [6] Mihai Surdeanu et al, Multi-instance Multi-label Learning for Relation Extraction. CoNLL 2012, 2012.

表1 各手法における正答率

手法	正答率
従来手法 [1]	0.4832
PV-DM	0.4969
ASV	0.4848
Weighted ASV	0.4866
提案手法	0.5021