

豊田工業大学 課題研究報告書要旨

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

企業がマーケティングのために行う商品の評判分析において、商品レビューに対するレーティング予測は重要な要素技術のひとつである。このレーティング予測では文字から文書に渡る様々な言語要素間の多様で複雑な関係を考慮する必要がある。しかし、複数カテゴリにおけるレーティング予測に関する従来手法 [1] は文間やカテゴリ間の関係を十分に考慮できていなかった。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の実現を目的とする。

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。藤谷ら [1] 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調整によって変化させカテゴリ間の関係性を考慮している。しかし、この手法は文同士の位置関係を考慮していない。また、カテゴリ間については考慮しているものの複雑な関係性を捉えられていない。

2.2 パラグラフベクトル

文や文書といった大きな単位の言語要素の分散表現を学習する手法 [2] 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している [2]。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。

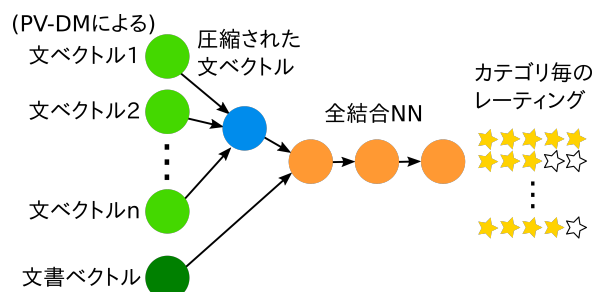


図 1: 提案手法による分類モデルの概略

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 (Neural Network, NN) を用いた評判分析の手法が、Nal ら [3], Rie ら [4], Duyu ら [5] 等に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 NN と全結合 NN 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分類問題を対象としている。つまり、多カテゴリの評判分類問題において、これらの手法により 1 カテゴリ毎に分類問題を解くだけではカテゴリ間の関係を考慮できない。

3 提案手法

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 NN の入力として分類を行う手法を提案する。文毎の意味表現を用いることで文同士の位置関係を考慮し、全結合 NN による分類器を用いることで文間及びカテゴリ間の複雑な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。

図 1 にモデルの概略を示す。提案手法は、まず、PV-DM によって各レビューとそれに含まれる文の意味表現となるベクトルを生成する。次に、各レビュー内の文ベクトルは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、レビュー毎に疎らだった文の数を統一する。最後に、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を分類器の入

力層として分類を行う。出力層の各ニューロンの出力値はあるカテゴリにおけるあるクラスの正規化されていない対数確率を表す。各カテゴリ内で対数確率が最も高いクラスがそのカテゴリの予測レーティングとなる。

4 実験

4.1 実験設定

提案手法が従来手法より優れていること、及び、文書ベクトルと文ベクトルを同時に用いることが評判分類に有効であることを示すため、各手法の正答率を測定した。比較手法として、提案手法における分類器の入力をそれぞれ (1) DV (Document Vector), (2) ASV (Averaged Sentence Vector), (3) Weighted ASV に変えた手法を用いた。DV とはレビュー全体の文書ベクトルであり、Quoc ら [2] の手法に相当する。ASV とはレビュー内で平均した文ベクトルであり、Weighted ASV と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である。また、データセットには先行研究 [1] と同一の楽天トラベルにおけるレビューデータを用いた。有意差検定にはマクネマー検定を用いた。

4.2 結果と考察

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の正答率を 0.0198 明らかに上回っていることから、提案手法が従来手法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 が ASV の正答率を 0.0029 有意に上回っている ($p < 0.05$) 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、提案手法が正答率において DV を 0.0050 有意に上回っており ($p < 0.001$), Weighted ASV を 0.0163 有意に

上回っている ($p < 0.001$) ことから、文書ベクトルと文ベクトルを同時に素性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。

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、提案手法が従来手法 [1] より高い正答率を示した。また、比較手法の結果より、レビュー内の文の並びが評判分類に重要であることが分かった。

今後の課題は言語要素間のより多様で複雑な関係を考慮することである。このためには、各レビューの意味表現を生成するモデルと分類を行うモデルを 1 つに統合する必要がある。なぜならば、モデルが分かれていることによって単語や文字などのより小さな言語要素同士の関係を分類時に考慮できないためである。モデルの統合によって、学習手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 et al., Distributed representations of sentences and documents. ICML 2014, 2014.
- [3] Nal Kalchbrenner et al., A convolutional neural network for modelling sentences. ACL 2014, 2014.
- [4] Rie Johnson et al., Effective use of word order for text categorization with convolutional neural networks. NAACL 2015, 2015.
- [5] Duyu Tang et al., 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
DV	0.4980
ASV	0.4838
Weighted ASV	0.4867
提案手法	0.5030