豊田工業大学 課題研究報告書

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

平成28年 2月

工学部 先端工学基礎学科
知能数理研究室
12056 外山洋太

目次

1　序論

企業において商品の評判分析のためのレビューの評判分類は重要な問題である。何万件という大量のレビューデータを人手で処理することは難しく、計算機による自動化が望まれる。その中で商品を複数のカテゴリにおいて分類をする問題がある。カテゴリとは、宿泊施設のレビューを例にすると、サービス、立地、食事等のレーティングが付けられる各項目のことである。この問題に関する従来手法 [1] は、文間の関係性を考慮しておらず、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。

近年、その評判分類において、ニューラルネットワークを用いた手法 [3, 4, 5] が提案されており、従来の手法を上回る正答率を達成している。ニューラルネットワークを分類問題に用いる利点はまず層の数を増やすことによって入力の深い繋がりを考慮できることである。例えば、文毎の素性を入力とすれば文間の関係性を捉えることができる。さらに、多カテゴリの分類問題においてはカテゴリ間の関係性を捉えた分類が実現できる。しかし、評判分類に関する多くの研究は 1 つのカテゴリにおける二値分類問題を対象としている。

文や文章の意味表現の学習手法として、単語と文章の分散表現を同時に学習するパラグラフベクトル [2] がある。これは評判分類問題に対して優れた性能を示している。しかし、文書全体にパラグラフベクトルを用いた場合、分類時に文の位置関係を考慮できない。

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。

提案手法では、パラグラフベクトル [2] によって生成された各レビューの文書ベクトルと文ベクトルをニューラルネットワークによる分類器において分類しレーティング予測を行う。楽天トラベルのデータセットを用いた実験において、提案手法は従来手法 [1] に対して約 2pp 上回る正答率を示した。レーティングの平均二乗誤差 (RMSE) を元にした評価基準では、従来手法 [1] において弱点となっていたカテゴリについてそれを上回る結果を示した。

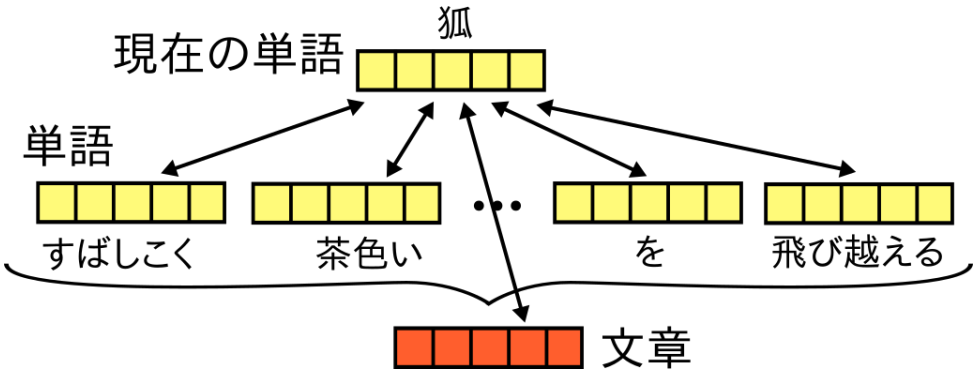

図 1: パラグラフベクトルの学習の概略

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) [7] モデルを用いた手法を提案している。その手法では、レビュー内の各文毎に予測した隠れ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。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。各文の素性には Bag Of Words (BOW) または Bag Of n-gram を用いている。各文毎に隠れレーティングを予測することによって 0.4832 の正答率が得られることが示された。また、カテゴリ間の繋がりによって正答率が変化することも示されている。

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意味表現を学習する手法である。これは、Continuous BOW (CBOW) または Skip-gram[6] という単語の意味表現の学習手法を応用した手法である。ここでは CBOW を応用した Distributed Memory model of Paragraph Vectors (PV-DM) について説明する。PV-DM は BOW と異なり、単語の並び順を考慮した文や文章の分散表現を生成することができる。

以下に具体的なアルゴリズムを示す。ここでは文章の意味表現を学習する場合について考える。学習の概略を図 1 に示す。まず、意味表現を学習する対象となる文章に含まれる単語を初めから一つず

つ読んでいく。その際、現在の単語及びその周辺の単語、現在の文章について、式 1 に示す目的関数 L を最大化するように各パラメータの学習を行う。

$$L = \frac{1}{T}\sum_{t=k}^{T} \log p(w_t|w_{t-k}, ..., w_{t-1}), \tag{1}$$
$$p(w_t|w_{t-k}, ..., w_{t-1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}},$$
$$y = b + Uh(w_{t-k}, ..., w_{t-1}, d; W, D)$$

ここで、d は文章、w_i は単語、W は全ての単語の分散表現を表す行列、D は全ての文章の分散表現を表す行列である。k はウィンドウサイズ、T は現在の文章に含まれる単語数である。ある単語の周辺を表す区間をウィンドウという。p は softmax 関数により正規化された、文脈から現在の単語が導かれることの尤度である。p を構成する y は現在の単語とウィンドウ内の単語及び現在の文章から導出される。$h(w_{t-k}, ..., w_{t-1}, d; W, D)$ は引数となるベクトルを平均したベクトルまたは結合したベクトルを返す関数である。

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 BOW 等に比べ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。まず、単語の意味表現から畳み込みニューラルネットワークを用いて単語同士の関係を捉えた特徴量を抽出する。その後、そこから得られた文書全体の特徴量を全結合ニューラルネットワークの入力とし多値または二値分類を行う。また、Duyu ら [5] と Nal ら [3] の手法はニューラルネットワークのモデルの中にパラメータとして単語の意味表現を取り込んでいる。これにより、特定の分類問題に対してそれを直に微調整することが可能となる。

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、パラグラフベクトルによってレビュー内の各文及び文章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。以下にその基礎となるアイデアと具体的なアルゴリズムを示す。

3.1　アイデア

先行研究 [1] の実験結果から、レビュー内の各文の特徴量を元にレビューの分類を行うことが分類精度の向上に有効であると考えられる。また、カテゴリ間の繋がりの変化が分類精度に影響していることから、これをパラメータとして機械学習のモデルに組み込めば分類精度を向上させることができると考えられる。

さらに、レビュー内の各文毎に分散表現を生成し分類器の入力とすることで、その順序を考慮した学習を行う。これにより、レビュー内の文の位置関係がレーティング予測時に利用できると考えられる。以下に、文の位置関係が重要となる例を示す。2 つ目の例は、1 つ目の例の 1 つ目の文と 3 つ目の文を入れ替えたものである。

> 食事は本格的で新鮮な食材が使われており美味しかった。しかし、それよりも嬉しかったことがある。それは、部屋からの海の眺めが素晴らしかったことである。これには子供も喜んでいた。

> それは、部屋からの海の眺めが素晴らしかったことである。しかし、それよりも嬉しかったことがある。食事は本格的で新鮮な食材が使われており美味しかった。これには子供も喜んでいた。

1 つ目の例では、4 つ目の文が直前の立地が良かったという文の意味を補完しているのに対し、2 つ目の例では、食事が良かったという文の意味を補完している。このように、文の位置関係によってどの文がどの文と強く関連しているかが変化する。それによって推定すべきレビュー全体のレーティングも変わると考えられる。よって、文の位置関係を考慮することは重要である。

しかし、個々のレビューに対して全ての文ベクトルを用いるのには、問題がある。なぜならば、各レビュー内の文の数はまちまちであり、データ上単純な行列として表せないためである。これは、ミニバッチ方式の訓練を難しくし、実行時間の増加に繋がる。この問題に対処するため、本手法では各レビュー内の文ベクトルに対して重み付け平均を行った。これにより、全てのレビューで文ベクトルの数が揃い、複数レビュー内の複数の文ベクトルをまとめて 3 次元行列として表すことができる。つ

まり、ミニバッチ方式の計算が容易となる。

分類器はニューラルネットワークを用いて構成することによって、文の位置関係とカテゴリ間の関係を同時に捉えた分類を行う。従来手法 [3] や [4]、[5] では、単語ベクトルに対して畳込みニューラルネットワークが用いられている。しかし、提案手法においては、畳み込みニューラルネットワークより全結合ニューラルネットワークを用いた方が精度が高かったため、分類器には後者のみを用いた。

3.2 アルゴリズム

提案手法の処理の流れを説明する。図 2 にアルゴリズム全体の概略を示す。

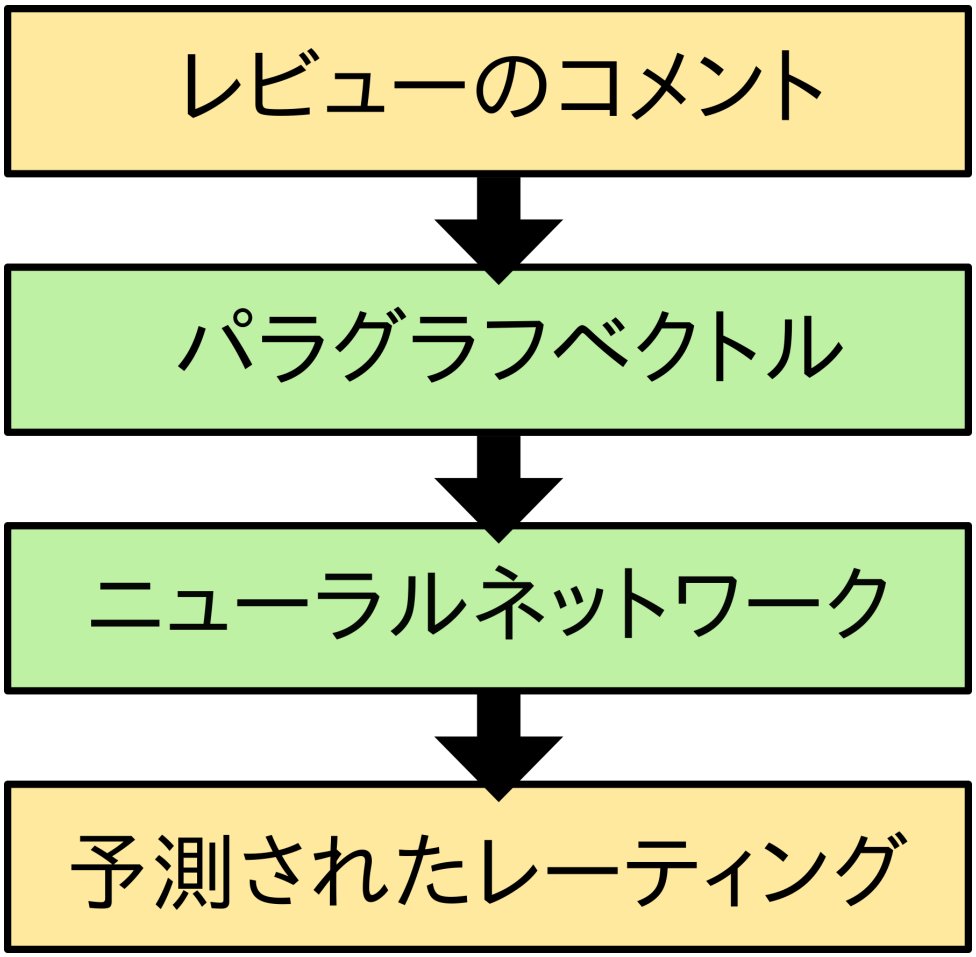

図 2: 提案手法におけるアルゴリズムの概略

初めに、PV-DM を用いて、各レビューの文書ベクトルとそれに含まれる各文のベクトルを生成する。文書ベクトルと文ベクトルについては別々に学習し生成する。式 1 の目的関数における h には引数のベクトルを結合する関数を用いる。また、学習の高速化のため、Quoc ら [2] によって用いられている階層的 softmax の代替として、ネガティブサンプリングを行う。ネガティブサンプリングとは、文脈外の単語をデータセットにおける出現確率でサンプリングし、それらと文脈の意味が遠ざか

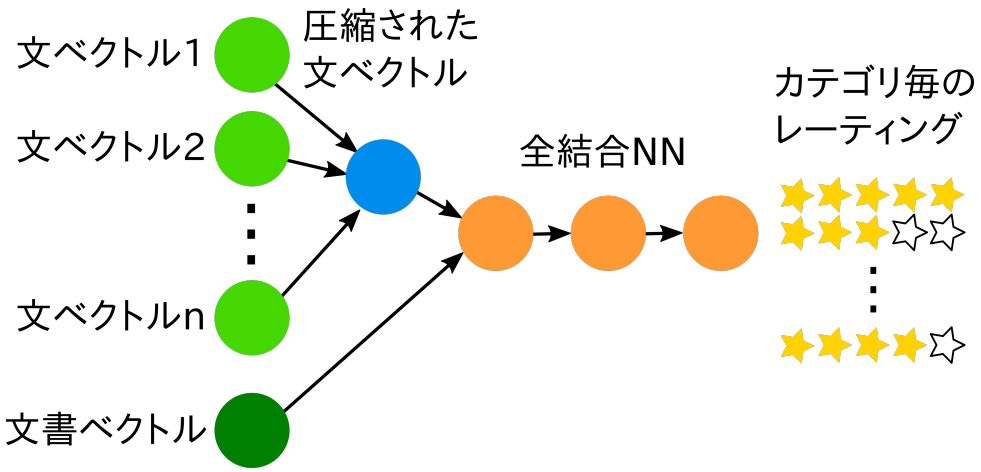

図 3: 全結合ニューラルネットワークによる分類器

るように学習する手法である。ただし、出現頻度は極端に頻出する単語の影響を抑えるため各単語に対して 3/4 乗している。現在の単語と同じ単語や同一回の学習で一度サンプリングされた単語はサンプリングしない。

次に、各レビュー内の全ての文ベクトルに対して重み付け平均を行い、圧縮された文ベクトルを生成する。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。式 2 に重み付け平均によって圧縮した文ベクトル $t_{i_{part}}$ を示す。各文ベクトルは圧縮後の各文ベクトルと位置が近いほど重みが大きくなるように重み付け平均する。

$$\mathbf{t}_{i_{part}} = \sum_{i_{sent}} \frac{w(x_{i_{part}}(i_{sent}))}{|\sum_{i'_{sent}} w(x_{i_{part}}(i'_{sent}))|} \mathbf{s}_{i_{sent}}, \tag{2}$$

$$x_{i_{part}}(i_{sent}) = \frac{i_{sent} - i_{part}}{\#partitions},$$

$$w(x) = \begin{cases} \frac{1}{2}(\cos(\pi|x|) + 1) & \text{if } |x| <= 1 \\ 0 & \text{otherwise} \end{cases}$$

ここで、i_{sent} はレビュー内の文のインデックス、$\#partitions$ は重み付け平均された後の文ベクトルの数、i_{part} は重み付け平均された後の文ベクトルのインデックス、$\mathbf{s}_{i_{sent}}$ は文ベクトルである。[*1]

最後に、分類器によってレーティング予測を行う。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 3 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。ニューラルネットワークの活性化関数には、シグモイド関数を用いる。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ち、各ユニットの出力はあ

[*1] 重み付けの関数には cos 関数の他に、x に対して線形に重みを減少させるような関数や、単純に文を区画毎に平均するような関数も考えられる。区画毎に平均する関数は他の 2 つより正答率が低く、線形な関数と cos 関数はほぼ同じ正答率を示したため、cos 関数を採用した。

るカテゴリ内であるクラスが選ばれることの正規化されていない対数確率を表す。ニューラルネットワークは式 3 に示す目的関数 E を最小化するように学習を行う。

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w), \tag{3}$$

$$y_{ck}(x_n; w) = \frac{e^{u_{ck}(x_n;w)}}{\sum_{j=1}^{K} e^{u_{cj}(x_n;w)}}$$

各ユニットの出力はカテゴリ毎に交差エントロピー誤差関数によって損失に変換される。ここで、u_{ck} は出力層のユニットの出力値、y_{ck} はカテゴリ c においてクラス k が選ばれる確率、w はニューラルネットワークのパラメータである。d_{nck} は n 番目の文書がカテゴリ c でクラス k ならば 1、それ以外で 0 となる値である。N はミニバッチサイズ、C はカテゴリの総数、K はクラスの総数である。

4　実験

4.1　実験設定

実験は、各手法の正答率を測定するものと、提案手法における予測レーティングと正解レーティングの平均二乗誤差 (RMSE) を測定するものの 2 つを行った。RMSE の計算において、正解または予測レーティングが 0 点であるものは評価から省いた。

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。これらの手法と提案手法に用いる分類器の入力を表 1 に列挙する。

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。

表 2 に各手法におけるニューラルネットワークのパラメータ設定を示す。全ての手法において、中間層の数は 1、入力層及び中間層におけるドロップアウト率はそれぞれ 0.2 と 0.5 で共通である。Weighted ASV と提案手法において圧縮された文ベクトルの数はそれぞれ 3 つと 2 つとした。全ての実験において文書及び文ベクトルについては、学習回数は 1,024 回、学習する単語の範囲は前 3 単語、単語の最少出現回数は 5 回、ネガティブサンプリングの回数は 5 回、ベクトルの次元数は 600 次元に設定し学習したものを用いた。

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 3 に示す。結果より、提案手法と 3 つの比較手法全てが従来手法の正答率を上回っている。提案手法が従来手法 [1] の正答

表 1: 各手法に用いられる特徴量

手法	特徴量
PV-DM	レビュー全体の文書ベクトル
ASV	レビュー内で平均した文ベクトル
Weighted ASV	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
提案手法	レビュー全体の文書ベクトル、 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

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 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。

次に、表 4 にレーティングの RMSE を測定した結果を示す。提案手法は従来手法 [1] が苦手としていた食事と風呂のカテゴリにおいてそれぞれ 0.58 及び 0.27 だけ低い誤差を示した。藤谷ら [1] より、本実験で用いたデータセットには、食事のカテゴリにおいて 0 点が付与されたレビューが 108,079 件存在する。また、風呂のカテゴリでは 13,332 件、設備のカテゴリでは 2,011 件存在し、他のカテゴリでは 0 件である。一般に、0 点はユーザが何らかの理由でレーティング不可能と判断したことを示す。よって、提案手法は従来手法 [1] よりレビュー中の上記のような意味をよく捉えていると考えられる。また、その他全てのカテゴリにおいても提案手法は従来手法より低い誤差を示した。

表 2: 各手法のパラメータ設定

手法	学習回数	中間層でのユニット数
PV-DM	20	512
ASV	55	256
Weighted ASV	24	256
提案手法	30	512

表 3: 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**

表 4: 提案手法と従来手法 [1] におけるレーティングの RMSE

手法	提案手法	従来手法 [1]
立地	0.88	0.97
部屋	0.90	0.97
食事	0.95	1.53
風呂	1.00	1.27
サービス	0.87	0.94
設備	0.93	0.95
総合	0.74	0.81

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。

謝辞

本研究において、楽天株式会社よりホテル予約サイト楽天トラベルにおけるレビューデータを使用させていただきました。この場を借りて深く感謝致します。

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] Yoshua Bengio et al, A Neural Probabilistic Language Model. The Journal of Machine Learning Research 3, 2003.

[7] Mihai Surdeanu et al, Multi-instance Multi-label Learning for Relation Extraction. CoNLL 2012, 2012.