文間及びカテゴリ間の関係性を捉えたレーティング予測

外山洋太　　三輪誠　　佐々木裕
豊田工業大学 知能数理研究室

{sd12056, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

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が提案されており、従来の手法を上回る分類精度を達成している。ニューラルネットワークを分類問題に用いることの利点の一つは層の数を増やすことによって入力となる特徴量の深い意味を捉えられることである。多カテゴリの分類問題に適用すれば、カテゴリ間の関係性を捉えた分類が実現できる。さらに、文毎の特徴量を入力とすれば文間の関係性も捉えることができる。

文や文章の特徴量としては、パラグラフベクトル [2] が分類問題に対して優れていることが示されている。

以上より、本研究では、複数カテゴリにおける評判分類について、パラグラフベクトルと CNN を用いて文間及びカテゴリ間の関係性を捉えた分類を実現し、従来手法から分類精度を向上させることを目的とした。

提案手法は、パラグラフベクトルによって生成された各レビューの特徴量をニューラルネットワークの分類器において分類しレーティング予測を行う。特徴量としては、各レビューの文書ベクトルに加え、各レビュー内の文ベクトルの重み付け平均を用いた。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ており、文間及びカテゴリ間の関係性を捉えた分類を行う。

実験において、提案手法は従来手法 [1] に対して 2pp 以上上回る精度を示した。正解ラベルと予測したラベルの差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。ただし、いくつかのカテゴリでは従来手法と変わらないか、より悪い結果が示された。また、実験により、パラグラフベクトルが文書と文について適用された場合、文書ベクトルと文ベクトルは僅かに異なる特徴を捉えていることが示された。

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、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手法を提案している。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。各文の組成は BOW または n-gram であり、それらの順序は無視されている。

実験結果より、各文毎に隠れレーティングを予測することによって分類精度が向上すること、また、カテゴリ間の繋がりによって分類精度が変化することが示されている。

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 Bag Of Words (CBOW) または Skip-gram 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 PV-DM について説明する。

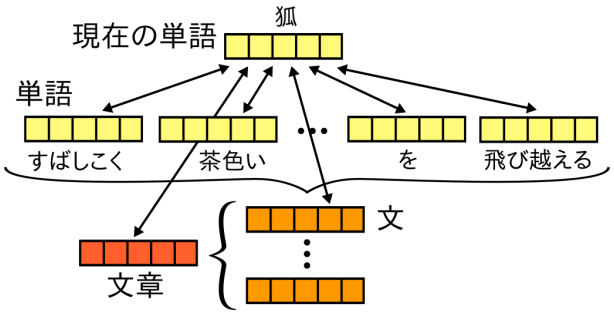

図 1: 文書及び文ベクトルの学習の概略

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、以下の式 1 に示す目的関数を最大化するように各パラメータの学習を行う。

$$L = \frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k},...,w_{t+k}) \tag{1}$$

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2}$$

$$y = b + Uh(w_{t-k},...,w_{t+k},d;W,D) \tag{3}$$

ここで、d は文、w_i は単語、W は全ての単語の分散表現を表す行列、D は学習している全ての文章の分散表現を表す行列である。 k は片側のウィンドウサイズ、T は現在の文章に含まれる単語数である。y は現在の単語とウィンドウ内のその周りの単語及び現在の文章から導出される正規化されていない対数確率である。 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。$h(w_{t-k},...,w_{t+k},d;W,D)$ 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。

また、学習効率を高めるために、ネガティブサンプリングを行う場合がある、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るように y の式を変更し学習する手法である。

これにより、Bag Of Words (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。

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、単語の分散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、単語の分散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あるいは文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、これら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分散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らを微調整することが可能になる。

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 アルゴリズム

実際のアルゴリズム全体の流れを説明する。

初めに、パラグラフベクトルを用いて、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のモデルを学習させ生成する。以下の目的関数を最大化するように学習を行う。

$$\begin{aligned} L_d = \sum_{t=1}^{T}\{&\log\sigma(s(w_t,w_{t-n},...,w_{t-1},d)) \\ &+ \sum_{w'_t\sim P_n}^{k} \log(1-\sigma(s(w'_t,w_{t-n},...,w_{t-1},d)))\} \end{aligned} \tag{4}$$

$$s(w_t, w_0, ..., w_n, d) = W_{score}(w_t) \cdot \begin{bmatrix} W(w_0) \\ \vdots \\ W(w_n) \\ D(d) \end{bmatrix} \quad (5)$$

ここで、T は現在の文書内の単語数、t は現在の単語位置、d は現在の文書、w_i は位置 i にある単語である。$W(w_i)$ は単語 w_i に相当する単語ベクトルを単語行列 W から抜き出す関数を表す。$D(d)$ は文書 d に相当する文書ベクトルを文書行列 D から抜き出す関数を表す。関数 $s(w_t, w_0, ..., w_n, d)$ はある単語とそれが出現する文脈との類似度を計算する。行列 W_{score} は。内積によって文脈と単語との類似度を計算するための単語毎のベクトルを保持する。文書行列内の各文書ベクトルはレビュー全体を表す文書ベクトル、または、書くレビュー内の文ベクトルを表す。また、式 4 の中括弧内の右項はネガティブサンプリングを表す。$w'_t \sim P_n$ は文脈外の単語 w'_t を単語の出現頻度 P_n によってサンプリングすることを示す。σ はシグモイド関数である。

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が統一される。以下に重み付け平均を行う関数を示す。

$$t_{i_{sect}} = \sum_{i_{sent}} \frac{\mathbf{w}(x_{i_{sect}}(i_{sent})}{|\mathbf{w}(x_{i_{sect}}(i_{sent}))|} s_{i_{sent}} \quad (6)$$

$$x_{i_{sect}}(i_{sent}) = \frac{i_{sent} - i_{sect}}{\#sections} \quad (7)$$

$$\mathbf{w}(x) = (w_1(x), w_2(x), ..., w_{i_{sect}}(x), ..., w_{\#sentences}(x)) \quad (8)$$

$$w(x) = \frac{1}{2}(\cos(\pi|x|) + 1) \qquad \text{if } |x| <= 1 \quad (9)$$

ここで、i_{sent} はレビュー内の文のインデックス、$\#sections$ は重み付け平均された後の文ベクトルの数、i_{sect} は重み付け平均された後の文ベクトルのインデックスである。$s_{i_{sent}}$ は文ベクトル、$t_{i_{sect}}$ は圧縮された後の文ベクトルである。

次に、レビュー毎の特徴量を用いて分類器によってレーティング予測を行う。分類器に用いる各レビューの特徴量としては、生成したレビュー全体の文書ベクトル及び圧縮された文ベクトルを用いる。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 2 に各層の結合の様子を示す。

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た。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それが含まれるカテゴリ内でそれが示すラベル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の最小化すべき目的関数を式 10 に示す。

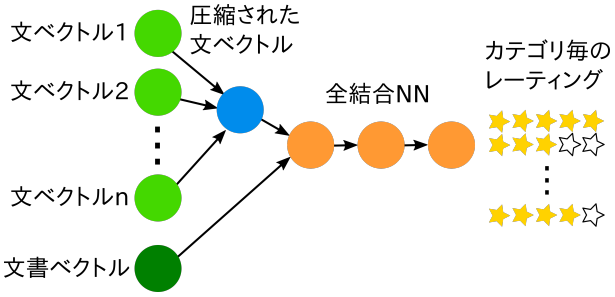

図 2: 全結合ニューラルネットワークによる分類器

表 1: 各基準手法に用いられる特徴量

手法	特徴量
基準手法	レビュー全体の文書ベクトル
提案手法 1	レビュー内で平均した文ベクトル
提案手法 2	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 3	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

$$E = -\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{k=1}^{K} d_{nck} \log y_{ck}(x_n; w) \quad (10)$$

$$y_{ck}(x_n; w) = \frac{e^{u_k}}{\sum_j^K e^{u_j}} \quad (11)$$

各ユニットの出力はカテゴリ毎にソフトマックスクロスエントロピー関数によって損失に変換される。ここで、$y_ck(x_n; w)$ はカテゴリ c においてクラス k が選ばれる対数確率、

提案手法に加え、基準手法として NN への入力となる特徴量が異なる 4 つの手法を用いた。各手法に用いた特徴量を表 1 に列挙する。

提案手法 1 と提案手法 2 の比較によって、文の位置関係が分類に対して重要であるかが示される。提案手法 2 と提案手法 3 の基準手法の比較によって、文書ベクトルに加え文ベクトルを用いることが有効であるかが示される。

表 2: 基準手法のパラメータ設定

項目	値
学習する単語の範囲	前 3 単語
単語の最少出現回数	5
ベクトルの次元数	600
分類器の学習回数	20
中間層の数	1
中間層でのニューロン数	512
入力層及び中間層でのドロップアウト率	0.2, 0.5

表 3: 提案手法 1 のパラメータ設定

項目	値
学習する単語の範囲	前 3 単語
単語の最少出現回数	5
ベクトルの次元数	600
分類器の学習回数	55
中間層の数	1
中間層でのニューロン数	256
入力層及び中間層でのドロップアウト率	0.2, 0.5

4 実験

基準手法及び提案手法について分類精度を測定するために実験を行った。

4.1 実験設定

実験は、分類精度を測定するものと、その実験で最も精度の高かった提案手法における予測レーティングと回答レーティングの平均二乗誤差を測定するものの 2 つを行った。分類精度を測定する実験では、1 つの基準手法及び 3 つの提案手法をそれぞれ同じデータセットに適用し、各手法の精度を測定した。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。レーティングの平均二乗誤差を測定する実験は、上記の分類精度を測定する実験で得られた予測レーティングを用いて測定を行った。表 2 と表 3、表 4、表 5 に各手法でのパラメータ設定を示す。

4.2 結果

表 6 に分類精度を測定する実験の実験結果及び従来手法による結果を示す。

表 4: 提案手法 2 のパラメータ設定

項目	値
学習する単語の範囲	前 3 単語
単語の最少出現回数	5
ベクトルの次元数	600
分類器の学習回数	24
中間層の数	1
中間層でのニューロン数	256
入力層及び中間層でのドロップアウト率	0.2, 0.5

表 5: 提案手法 3 のパラメータ設定

項目	値
学習する単語の範囲	前 3 単語
単語の最少出現回数	5
ベクトルの次元数	600
分類器の学習回数	30
中間層の数	1
中間層でのニューロン数	512
入力層及び中間層でのドロップアウト率	0.2, 0.5

表 6: 各手法における正答率

手法	精度
従来手法	0.4832
基準手法	0.4969
提案手法 1	0.4848
提案手法 2	0.4866
提案手法 3	0.5038

4.3 考察

表 6 より、基準手法が従来手法の正答率を上回っていることから、単語

5 結論

本研究では、レビュー全体の文書の分散表現に加え、レビュー内の各文ベクトルに対する分散表現の重み付き平均を用いた評判分類の手法を提案した。

実験により、従来手法 [1] 及びレビュー全体の文書ベクトルのみを用いた手法に比べ、提案手法が高い分類精度を示すことが分かった。同時に、これはレビュー内の文の並びが評判分類に重要であることを示す。

5.1 今後の課題

今後の課題は、2 つに分かれているモデルの統一である。

提案手法は、分類すべき文書とそれが含む文の分散表現を生成する段階、及び、それらを用いて分類を行う段階の 2 つの段階に分かれている。このことは、問題を 2 つに分けることで個々の問題を単純にしているが、同時に一つずつ文書の分類を行うことを難しくしている。また、提案手法において、文書や文の分散表現を事前に生成するためのパラグラフベクトルの手法におけるパラメータは、実際には最大の分類精度を達成するため分類器のパラメータとして最適化されることが望ましい。

今後は、これらの問題を解決するために、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取り込む。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分類精度の向上を目指す。

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.