1S-04

単語拡張によるテキスト分類精度の改善と評価

鳥山 修平 † 世木 博久 †
† 名古屋工業大学大学院 情報工学専攻

1 はじめに

Web 上の twitter やブログなどごく短い文書に効果的な分類の必要性は増加している．テキスト分類精度を向上させる工夫は様々存在しており, 本報告では単語拡張と特徴選択の 2 つの方法を行う．単語拡張 (term expansion) は短い文書のスパースな情報を補い, 分類精度を向上させる [1, 2]．次に, 特徴選択 (feature selection) はテキスト中の単語を厳選することで, 分類の効率化, 精度を向上させる [3].

本研究では，これら 2 つの方法を組み合わせ, マルチクラスショートテキスト分類問題への効果を評価する．また, 最近提案された特徴選択法 [4] について, 実験によりその効果を検証し, 改善方法を提案する．

2 テキスト分類のための単語拡張

ショートテキスト分類では単語の共起関係や共通する文脈情報が乏しい場合がある．そこで, 単語拡張は単語文書行列 T (TDM) に対して, 単語間類似度行列 $S = (s_{ij})_{1 \leq i,j \leq M}$ (M は単語数) を生成し, その積を構成することで拡張された単語文書行列 $T^{TE} = TS$ を生成し, 単語情報を補う (図 1)．ここで, 単語間類似度行列 S の要素 s_{ij} は単語 t_i と t_j 間の類似度を表す値である．

本稿では, s_{ij} として Dice 係数に基づく Sim_D, Jaccard 係数に基づく Sim_J をそれぞれ以下のように与えた [5]:

$$Sim_D(t_i, t_j) = \frac{2|t'_i \cap t'_j|}{|t'_i| + |t'_j|}, \quad Sim_J(t_i, t_j) = \frac{|t'_i \cap t'_j|}{|t'_i \cup t'_j|}.$$

ここで, t' は単語 t が出現する文書の集合を表す．

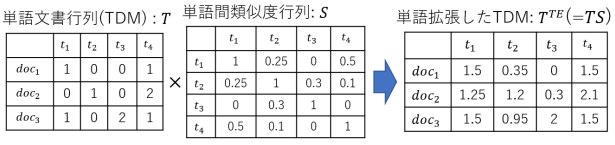

図 1: 単語拡張の例

表 1: 単語 t のクラス c に対する分割表

	c	$\bar{c}$	row
t が出現	tp	fp	N_t
t が出現しない	fn	tn	$N_{\bar{t}}$
col	pos	neg	N

3 テキスト分類のための特徴選択

文書分類を行う際, 小さなコーパスでも非常に多くの単語が含まれる場合がある．特徴選択は, それらすべての単語を扱うのではなく分類に有益な単語を厳選し, 分類の効率化を図る [3].

特徴選択は TDM と各文書のクラス情報を参照し分割表を生成する (表 1)．これをもとに各単語にスコアを算出し, スコアが高いものから順に選択する．[4] では, マルチクラスに対応した BNS (Bi-Normal Separation) を用いた特徴選択法 *EBNS* が提案された:

$$EBNS(t_i) = \max_k (F^{-1}(tpr_{c_k}(t_i)) - F^{-1}(fpr_{c_k}(t_i))). \quad (1)$$

ここで, c_k $(k = 1, ..., K)$ はクラスラベルを, F^{-1} は正規累積分布の逆関数を示し, tpr_{c_k} および fpr_{c_k} は以下の式で与えられる:

$$tpr_{c_k} = \frac{tp_{c_k}}{pos_{c_k}}, \quad fpr_{c_k} = \frac{fp_{c_k}}{neg_{c_k}}.$$

本稿では, *EBNS* に加え, ショートテキスト対応した特徴選択法 *S-EBNS* を新たに定義し, 従来法と比較した:

$$S\text{-}EBNS(t_i) = \max_k |F^{-1}(tpr_{c_k}(t_i)) - F^{-1}(fpr_{c_k}(t_i))|. \quad (2)$$

4 評価方法

本稿では, 2 つのデータセットを対象に実験を行った (表 2)．1 つ目は Reuters-21578 の中の単一ラベルをもつ R8 データセット†を用いた．本研究ではごく短いテキストの扱いに関心があるので, 各ニュースの見出し (title field) だけを入力文書とした．2 つ目は 20 Newsgroups††を対象とした．このコーパスでは, 1 文書に上限 30 単語となるようにショートテキスト化の処理を行った．

データの前処理として, 分類に不要なストップワードの削除, レンマ化を行った．テキスト分類のための分類器としては SVM (scikit-learn) を用いた．

Evaluation of Short Text Classification Using Term Expansion
† Shuhei Toriyama (31414093@stn.nitech.ac.jp)
† Hirohisa Seki (seki@nitech.ac.jp)
Dept. of Computer Science, Nagoya Institute of Technology (†)
Showa-ku, Nagoya, 466-8555 Japan

† https://ana.cachopo.org/datasets-for-single-label-text-categorization
†† https://scikit-learn.org/0.19/datasets/twenty-newsgroups.html

表 2: 実験対象のコーパス

(1) Reuters-21578 R8 データセット

Class	#(train)	#(test)	Total
acq	1604	688	2292
crude	261	113	374
earn	2746	1177	3923
grain	35	16	51
interest	190	82	272
money-fx	216	93	309
ship	100	44	144
trade	228	98	326
Total	5380	2311	7691

(2) 20 Newsgroups コーパスの部分データセット

Class	#(train)	#(test)	Total
rec.autos	730	185	915
rec.motorcycles	754	190	944
rec.sport.baseball	736	189	925
rec.sport.hockey	765	188	953
Total	2985	752	3737

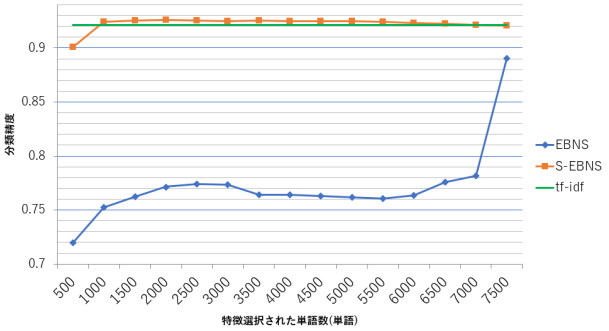

図 2: 単語数の推移による分類精度の変化 (Reuters)

5 実験結果

表 3 にそれぞれのデータセットにおけるテキスト分類精度の実験結果を示す. 表の上段の *tf-idf* は基準となる分類結果示しており, Sim_D は Dice 係数に基づく単語拡張を用いた分類結果を示す. 単語拡張では両データセットともに *tf-idf* との有意差が確認でき, 分類精度向上の効果が見られた.

次に, 特徴選択の実験結果について, 表 3 の (1), (2) はそれぞれ *EBNS*, *S-EBNS* の分類結果とその時の単語数を示す. 特徴選択の単語数は 500, 1000, 1500,... と変化させた. その時の単語数の変化による分類精度の推移を図 2 に示す. ここでは, 従来法 *EBNS* では精度が著しく低下し, データ圧縮の効果も確認できなかった. それに対し, 提案法 *S-EBNS* はデータセットによる精度向上の効果には差があるものの, 両データセットでも *tf-idf* と同程度の分類精度でデータ圧縮の効果が見られた.

表 3: 実験結果, (1),(2) は *EBNS*, *S-EBNS* (第 3 節) を表す. *は *tf-idf* との p 値 < 0.05.

(1) Reuters-21578 R8 データセット

	tf-idf	Sim_D	Sim_J	(1)	(2)
F_1	0.921	0.928*	0.925	0.890	0.925
単語数	7811	-	-	7500	1500

	(1)×Sim_D	(1)×Sim_J	(2)×Sim_D	(2)×Sim_J
F_1	0.895	0.890	**0.930***	0.928*
単語数	7500	7500	1500	1500

(2) 20 Newsgroups コーパスの部分データセット

	tf-idf	Sim_D	Sim_J	(1)	(2)
F_1	0.868	0.879*	**0.880***	0.848	0.862
単語数	5401	-	-	5000	4000

	(1)×Sim_D	(1)×Sim_J	(2)×Sim_D	(2)×Sim_J
F_1	0.858	0.856	0.866	0.871
単語数	5000	5000	4000	4000

最後に, 特徴選択の後に単語拡張を行った実験結果について, (1)×Sim_D は *EBNS* と Dice 係数を用いたものを示す. ここでは, 特徴選択で圧縮された入力に対しても単語拡張の分類精度向上の効果が確認できた.

6 まとめ

本研究では, ショートテキストの分類のための単語拡張と特徴選択を導入し効果を比較した. 単語拡張では分類精度向上の効果が確認できた. また, 特徴選択では精度向上には差があれど, データ圧縮の効果が見られた. 今後の課題として, 他のコーパスによる実験や他の特徴選択法との比較と評価が挙げられる.

参考文献

[1] Carpineto et al.: A Concept Lattice-Based Kernel for SVM Text Classification, *Formal Concept Analysis* (Ferré, S. and Rudolph, S., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 237–250 (2009).

[2] Boutari, A. M. et al.: Evaluating Term Concept Association Mesaures for Short Text Expansion: Two Case Studies of Classification and Clustering, *CLA* (2010).

[3] Forman, G. et al.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification, The Journal of Machine Learning Research (2003).

[4] Baillargeon, J.-T. et al.: Weighting Words Using Binormal Separation for Text Classification Tasks with Multiple Classes, Canadian Conference on Artificial Intelligence (2019).

[5] Seki, H. and Toriyama, S.: On Term Similarity Measures for Short Text Classification, IEEE 11th International Workshop on Computational Intelligence and Applications (2019).