分野タグの誤り検出と修正

長倉 和也 † **福本 文代** ‡

山梨大学工学部

g04mk018@ccn.yamanashi.ac.jp†, fukumoto@yamanashi.ac.jp‡

1 はじめに

近年, 共通のベンチマークとして Reuter-21578 や毎日新聞などのタグ付きデータが作成されるに伴い, 教師付き学習を用いた文書の自動分類に関する研究が多く行われている. 一般に教師付き学習の精度は, タグ付きコーパスの質に依存する. しかし, タグ付きコーパスは人手により作成されるため, 誤りが含まれている. 従って, 高精度で分類を行うためには, 分類対象となるコーパスに付与されたタグの誤りを自動的に検出し修正する技術が必要となる.

本研究では, 文書に付与された分野名の誤りに注目し, これを自動的に検出・修正する手法を提案する. 我々は, 誤り検出と修正の手がかりとして 2 つの点に注目した. 1 点目は対象となる全ての事例から誤りの候補を抽出するために高速な機械学習法の 1 つである Naïve Bayes(NB) を利用する点である. NB を用いて対象となる事例を学習・分類した結果, 予め付与されている分野名と異なる分野に分類された事例は誤りである可能性が高いと考える. 2 点目は候補の中から誤り事例を検出し修正するために損失関数を利用する点である. 損失は, テスト事例の真の分布と実際に機械学習を用いて求めた分布との差を示す. 誤り候補となる事例を訓練事例に加えたデータを用いて損失を求め, 損失の値により誤りであるか否かを判定した. 実験では RWCP コーパス [1] を用いて本手法の有効性を検証した.

2 損失を用いた誤り検出と修正

2.1 素性選択

誤り推定に用いる NB の分類精度を向上させるために素性選択を行った [2]. 素性選択法として Information Gain(IG) を用いた. IG は式 (1) で示される [3].

$$
\begin{aligned}
G(t) = &-\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) \\
&+P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) \\
&+Pr(\bar{t})\sum_{i=1}^{m} P(c_j|\bar{t})\log P_r(c_i|\bar{t})
\end{aligned} \tag{1}
$$

式 (1) において c_i は分野, t は訓練記事集合に出現する単語を示す.

2.2 誤り候補の検出

対象となる全ての事例から誤りの候補を抽出するために NB を用いて対象となる事例を学習・分類した. その結果, 予め付与されている分野名と異なる分野に分類された場合, この事例は誤りである可能性が高いとし, これらの事例を誤りの候補として抽出した. NB はいくつかの手法が提案されている [4]. 我々は, McCallum らによって提案された NB を用いた [5].

$$
P(c_j|d_i,\hat{\theta}) = \frac{P(c_j|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{d_{ik}}|c_j,\hat{\theta})}{\sum_{r=1}^{|C|} P(c_r|\hat{\theta})\prod_{k=1}^{|d_i|} P(w_{d_{ik}}|c_j,\hat{\theta})} \tag{2}
$$

式 (2) はテスト文書 d_i が分野 c_j に分類される確率を示す. $|C|$, 及び $|d_i|$ はそれぞれ分野数, 文書中における単語の異なり数を示す. $w_{d_{ik}}$ は文書 d_i 中に出現する単語 k を示す. 式 (2) において, $P(w_t|c_j,\hat{\theta})$, 及び $P(c_j|\hat{\theta})$ はそれぞれ式 (3), 及び (4) で示される.

$$
P(w_t|c_j,\hat{\theta}) = \frac{1+\sum_{i=1}^{|D|} N(w_t,d_i)P(c_j|d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(w_s,d_i)P(c_j|d_i)} \tag{3}
$$

$$
P(c_j|\hat{\theta}) = \sum_{i=1}^{|D|} P(c_j|d_i)/|D| \tag{4}
$$

$|V|$ は, 訓練データ中に出現する単語全体の集合を示し, $|D|$ は, 訓練データの個数を示す. また, $N(w_t,d_i)$ は

文書 d_i に出現する単語 w_t の個数を示し, $P(c_j|d_i)$ は, $P(c_j|d_i) \in \{0,1\}$ とする.

2.3 損失

損失とは機械学習を用いてテスト記事を分類した結果得られる分野と, 実際にテスト記事が分類されるべき分野との間に生じるずれである. すなわち, ずれが小さいほど損失は小さくなり優れた学習法であると言える. 予め分野名が付与された訓練記事集合 D を用いてテスト記事 $\mathbf{x}$ を分類した結果, 得られる分野名を y とすると, D に $(\mathbf{x}, y)$ を加えた新しい記事集合 D^* における損失は損失関数を用いることで求めることができる. 我々は, 2 種類の関数, すなわち log loss 法と 0/1 loss 法を用い検出と修正の精度を比較した. log loss 法を式 (5) に, 0/1 loss 法を式 (6) に示す [6].

$$\tilde{E}_{\hat{P}_{D^*}} = -\frac{1}{|\mathcal{P}|}\sum_{\mathbf{x}\in\mathcal{P}}\sum_{y\in\mathcal{Y}}\hat{P}_D(y|\mathbf{x})\log(\hat{P}_{D^*}(y|\mathbf{x})) \quad (5)$$

$$\tilde{E}_{\hat{P}_{D^*}} = \frac{1}{|\mathcal{P}|}\sum_{\mathbf{x}\in\mathcal{P}}(1-\max_{y\in\mathcal{Y}}(\hat{P}_{D^*}(y|\mathbf{x})) \quad (6)$$

式 (5), 及び式 (6) において, $|\mathcal{P}|$ は誤り推定に用いたデータの個数を示し, $\hat{P}_{\mathcal{D}}(y|\mathbf{x})$ は D において任意の $\mathbf{x}$ が分野 y に属すべき確率分布（真の分布）である. 真の分布はサンプリング予測により求めた. 文書に予め付与されていた分野 y_{old} と NB により分類された分野 y_{new} において式 (5), あるいは式 (6) を用いて損失を求める. 損失値の大小比較をした結果, y_{old} の損失値が y_{new} の損失値より大きい場合, 誤りであると見なし, 分野名を y_{new} へ修正する.

2.4 検出と修正法

本手法ではテスト記事集合から誤り候補を抽出し損失を求める. 検出・修正の処理を図 1 に示す.

1. 誤り候補の抽出
   訓練記事集合 D を用いて NB で学習し, テスト記事集合 T を分類する. 分類に失敗したデータの集合 $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ を誤り候補とする. X に予め付与されていた分野を y_{old}, NB で分類された分野を y_{new} とする.
2. 損失の推定
   T から X を除いたデータを T^* とする. 誤り候補となるデータ $\mathbf{x}_i (1 \le i \le n)$ 全てに対して以下の作業を行う.

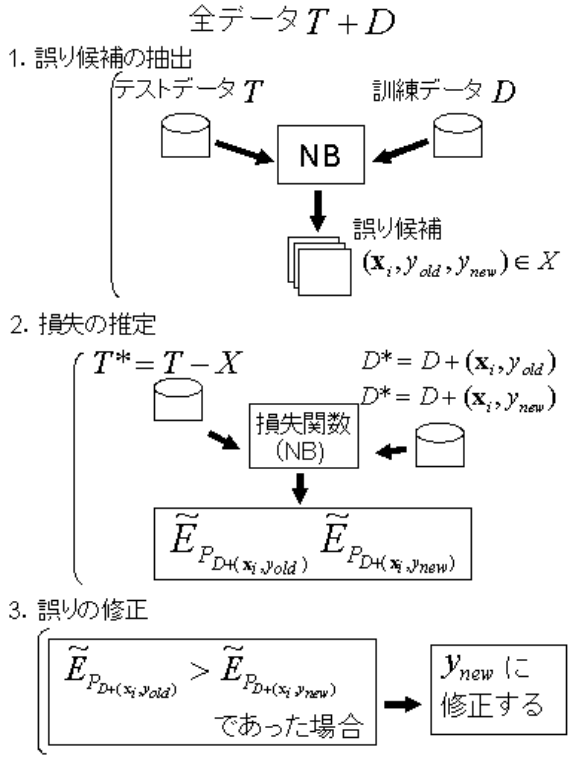

図 1　検出・修正の処理

   (a) $D + (\mathbf{x}_i, y_{old})$ を訓練記事集合として NB で学習し, テスト記事集合 T^* を分類する.
   (b) (a) の結果と損失関数を用いて $\tilde{E}_{P_{D+(\mathbf{x}_i, y_{old})}}$ を求める.
   (c) $D + (\mathbf{x}_i, y_{new})$ に対して (a), 及び (b) の処理を用いて $\tilde{E}_{P_{D+(\mathbf{x}_i, y_{new})}}$ を求める.
3. 誤りの修正
   $\tilde{E}_{P_{D+(\mathbf{x}_i, y_{old})}} > \tilde{E}_{P_{D+(\mathbf{x}_i, y_{new})}}$ ならば y_{old} は誤りであると見なし y_{new} に修正する. そうではない場合は y_{old} が正しいと見なし修正を行わない.

対象とするデータの誤り検出と修正の精度は 3 回の交差検定により求めた.

3 実験

実験データとして, RWCP コーパス毎日新聞 1994 年の 10,195 記事を使用した [1]. これらの記事は, 国際十進分類の中の政府, 国際, 犯罪, 軍事, 教育システム, 農林水産, 交通, 演劇, スポーツの 9 分野のいずれかに属している. 各分野の記事数を表 1 に示す. データは形態素

表 1　誤りの検出と修正の精度

分野名	記事数	分野名	記事数
政府	1211	農林水産	442
国際	1285	交通	783
犯罪	1963	演劇	457
軍事	337	スポーツ	3331
教育システム	386		
		合計	10195

表 2　誤りの検出と修正の精度

	Log 損失関数	0/1 損失関数
検出数	569	391
検出成功	**56.9%** (324)	52.7% (206)
修正成功	52.2% (297)	49.9% (195)
修正/検出	91.2%	94.7%

解析を用いて文書に出現するサ変名詞, 一般名詞, 固有名詞を抽出したものを使用した.

3.1　誤り検出と修正

誤り検出と修正を行った実験結果を表 2 に示す. 表 2 から log loss 法と 0/1 loss 法の精度を比較すると, log loss 法の方が検出数も多く, 検出成功率, 修正成功率ともに高いことがわかる. 検出が成功した中で修正も成功している記事は, 9 割以上と高い結果となった. しかし, 検出の精度は 56.9% であった. よって, 検出成功率を向上させることが今後の課題となる.

表 3 は検出と修正を行った例である. ○は正しい分野名を表し, ×は誤った分野名を示す. 表 3 において, ラグビーの選手名鑑がスポーツから犯罪に修正されており, 修正が失敗している. この記事の内容を見てみると背番号と選手名が載っていた. 選手名は形態素解析されると名字と名前に分けられているため, 名字が同じである人が, 他の分野を特徴づける重要な単語となっている場合, 他の分野に修正されてしまっていると考えられる. 実際にこの記事では「中川」という選手が載っていた. 用いた記事の中で中川知事の闇献金疑惑という内容の記事が多く出現したため, その影響を受け誤って犯罪に分類されたと考えられる.

3.2　繰り返しによる誤り修正

誤り検出された記事が正しく修正されたと考え, 新しく付与された分野を元のデータに反映させ繰り返し誤り検出と修正を行った. 誤り検出数の推移を図 2 に, 文書分類の精度を図 3 に示す. 図 2, 及び図 3 において横軸は繰り返し回数を示す. また, 図 2 の縦軸は誤り検出された記事数, 図 3 の縦軸は文書分類の精度を示す.

表 3　誤りの検出と修正例

該当記事のタイトル	修正前	修正後
細川首相の年頭会見の要旨	犯罪×	政府○
全国高校ラグビー 出場校選手名鑑 広島工	スポーツ○	犯罪×
ペルー大統領選への出馬は本気	犯罪×	政府×

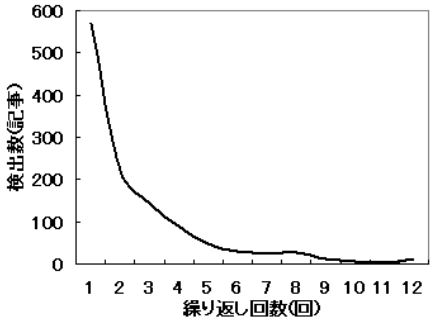

図 2　誤り検出数の推移

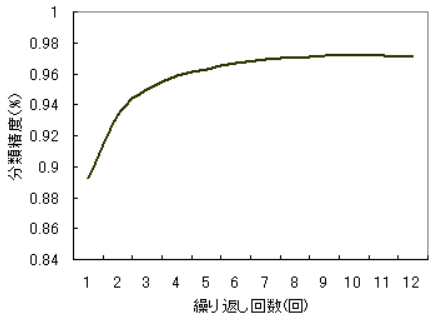

図 3　分類精度の推移

図 2, 及び図 3 から誤り修正を繰り返すことで誤り検出数が少なくなり分類精度が向上しているのがわかる. しかし, 図 2 において繰り返し回数が 7 回から 8 回で検出数は 23 記事から 26 記事となっていることから, 繰り返し回数を増やすことで検出数が増える場合も存在した. さらに, 11 回から 12 回で 4 記事から 12 記事となっていた. これは, 7 回目の修正では '政府' となり 8 回目の修正では '国際' であると判断された記事が多く出現し, 誤り検出と修正を繰り返すことで '政府' と '国際' が交互に付与されたためであると考えられる. 各回における損失値を求めた結果, $\tilde{E}_{P_{D+(\mathbf{x}_i, y_{\text{政府}})}} \fallingdotseq \tilde{E}_{P_{D+(\mathbf{x}_i, y_{\text{国際}})}}$ と

表 4　人工誤り記事に対する誤り検出と修正の精度

加えた誤り	5% (510)	10% (1020)
検出成功	55.9% (285)	53.3% (544)
修正成功	50.2% (256)	46.4% (473)
修正/検出	89.9%	86.9%
加えた誤り	20% (2039)	30% (3059)
検出成功	48.4% (987)	44.3% (1355)
修正成功	43% (877)	38.8% (1186)
修正/検出	88.9%	87.5%

表 5　誤り候補に含まれる誤りを付与した記事数

加えた誤り	5% (510)	10% (1020)
誤り候補中の人工誤り記事数	507	1006
加えた誤り	20% (2039)	30% (3059)
誤り候補中の人工誤り記事数	2010	3018

なっていたため, 例えば, 7 回目で '国際' に修正された記事数が多い場合, 8 回目の修正では, '国際' に修正された記事の影響を受け '国際' に修正されてしまったと考えられる.

3.3　人工的に誤りを付与

人工的に誤りを付与した記事について誤り検出と修正を行った. 本実験では, 誤りのデータ数を全データの 5%, 10%, 20%, 30% の各場合について実験を行った. 実験結果を表 4 に示す. 表 4 において, 検出成功は人工的に誤りを付与したデータのうち検出が成功した割合, 修正成功は人工的に誤りを付与したデータのうち検出と修正が成功した割合を表し, 修正/検出は検出に成功した記事の中で修正も成功した割合を表している. また, 括弧内の数値は記事数を表している. 表 4 から誤りを付与した記事の割合が高くなるほど, 誤り検出の精度が低下していることがわかる. また, 検出の精度が 45%〜55% であり良好でないこともわかる. 精度が上がらなかった原因として 2 点考えられる. 1 点目は誤り候補を抽出するために用いた NB の精度. 2 点目は誤りの検出と修正を行うために用いた損失関数である. 人工的に誤りを付与した記事がどの程度, 誤り候補に選ばれているかを調べた結果を表 5 に示す. 表 5 から, 例えば誤りを付与した割合が 5% の時, 510 記事のうち 507 記事が抽出されていることから, 誤り候補として人工的に誤りを付与した記事のほとんどが抽出されていることがわかる. このことから, 損失関数に問題があると考えられる. 本手法では損失値 $\tilde{E}_{P_{D+(\mathbf{x}_i,y_{old})}}$ と $\tilde{E}_{P_{D+(\mathbf{x}_i,y_{new})}}$ の大小関係のみから誤りであるか否かを判定していた. 従って, 誤りでないにもかかわらず修正後の損失値が修正前よりも小さい場合には誤りであると判定されてしまう. 今後は効果的な閾値の設定方法を検討する必要がある.

4　まとめ

本稿では, 文書に付与された分野名の誤りに注目し, これを自動的に検出・修正する手法を提案した. RWCP コーパスを用いた実験の結果, 検出した記事については修正精度が 91.2% であり, 修正の有効性が確認できた. 一方, 誤り検出の精度は 52.2% であったことから, 今後の課題としては, さらに検出の精度を向上させるため, 効果的な閾値の設定方法や複数のデータを用いた多数決方式による検出方法について検討する必要がある. また, 大規模コーパスを用いた定量的な評価を行う必要がある.

参考文献

[1] RWC: *RWC Text Database (Japanese)*, Real World Computing (1995).

[2] Yiming Yang, Xin Liu: A re-examination of text categorization methods, *Proc. of the 22nd Annual International ACM SIGIR Conference on Reseach and Development in Information Retrieval*, pp. 42-49 (1999).

[3] Yiming Yang, Jan O. Pedersen: A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the 14th International Conference on Machine Learning (ICML '97)* pp. 412-420 (1997).

[4] Tom Mitchell: *Machine Learning*, McGraw Hill (1996).

[5] Andrew McCallum, Ronald Rosenfeld, Tom Mitchell, Andrew Y. Ng: Improving Text Classification by Shrinkage in a Hierarchy, *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, pp. 359-367 (1998).

[6] Nicholas Roy, Andrew McCallum: Toward Optimal Active Learning through Sampling Estimation of Error Reduction, *Proceedings of the 18th International Conference on Machine Learning (ICML '01)* , pp. 441-448 (2001).