6C-5

文脈木重み付け法を用いた半教師付き学習による文書分類の検討

小畑智広† 小林 学‡ 渡辺重佳‡

†湘南工科大学大学院
‡湘南工科大学 工学部 情報工学科

1. はじめに

与えられたカテゴリが既知のデータ(学習データ)を用いて, カテゴリが未知のデータ(新規データ)がどのカテゴリに所属するかを自動的に判別する自動分類問題は, ベクトル空間モデルやサポートベクターマシン[1](以下SVMと略す)などの手法により大きく発展した. 一方, J.Zivらは無ひずみデータ圧縮法であるLZアルゴリズムを用いて文書分類を行う手法を提案した[2]. さらに学習文書及び新規文書の長さが無限に近づくときには, 正しい分類が行えることを示した[3]. またLZアルゴリズムよりも圧縮性能の優れた文脈木重み付け法[4〜6](以下CTW法と略す)を文書分類に適用する手法も提案されている[6, 7]. この手法は各カテゴリが同一の確率モデルから生起することを仮定している. 一方, 文書分類の場合, 1カテゴリ中に複数のトピックが混在している場合も考えられる. そこで本研究では, 各カテゴリ中に複数の確率モデルが混在する場合を想定し, サブカテゴリを構築することを検討する. 具体的には少数の学習文書に対してサブカテゴリを手動で構築し, その後全学習文書に対してCTW法を用いた半教師付き学習法[8]を適用する. また新聞データに本手法を用いて計算機実験を行い, 正分類率によりその有効性を示す.

2. CTW 法を用いた文書分類

本節では CTW 法を用いて文書分類を行う手法[7]を述べる. 今ある 2 元入力系列 $\boldsymbol{x}$ に対する CTW 法の出力(重み付け確率)を$P_w^{\lambda}(\boldsymbol{x})$ と表記する. また理想符号長 $L(\boldsymbol{x})$ を式(1)で定義する.

$$L(\boldsymbol{x}) = -\log_2 P_w^{\lambda}(\boldsymbol{x}). \quad (1)$$

なお$L(\boldsymbol{x})$は圧縮後のビット数の理論値に対応している. 次に文書の自動分類において, 分類カテゴリは 1〜C まで存在するものとする. ここでカテゴリ $i \in \{1,2,\cdots,C\}$ に所属するN_i 個の学習文書をそれぞれ$\tilde{x}_1^{(i)}, \tilde{x}_2^{(i)}, \cdots, \tilde{x}_{N_i}^{(i)}$ と書く. ただし各文書の文書長が十分に大きくない場合を想定し, 各カテゴリに対する文書が同一の確率モデルから生起するという仮定を置く. ここでカテゴリ i に所属する全ての学習文書を一つの学習データとして連結したものを$\tilde{x}^{(i)} = \tilde{x}_1^{(i)}\tilde{x}_2^{(i)}\cdots\tilde{x}_{N_i}^{(i)}$ と表記する. このとき上の仮定から, $\tilde{x}^{(i)}$に対して CTW 法によりカテゴリ i の文脈木を構築する. これは各カテゴリに対して 1 回だけ行えばよく, 結果の文脈木を保存しておくものとする. また CTW 法を用いることにより全ての $i \in \{1,2,\cdots,C\}$ に対して$P_w^{\lambda}(\tilde{x}^{(i)})$を計算し$L(\tilde{x}^{(i)}) = -\log_2 P_w^{\lambda}(\tilde{x}^{(i)})$ が得られる.

次にカテゴリが未知の新規文書を $\boldsymbol{x}$ とし, $P_w^{\lambda}(\tilde{x}^{(i)})$ の続きから新規文書$\boldsymbol{x}$ を結合して, $P_w^{\lambda}(\tilde{x}^{(i)}\boldsymbol{x})$を新たに計算する. このとき$\tilde{x}^{(i)}\boldsymbol{x}$ に対する理想符号長は$L(\tilde{x}^{(i)}\boldsymbol{x}) = -\log_2 P_w^{\lambda}(\tilde{x}^{(i)}\boldsymbol{x})$となる. $L(\tilde{x}^{(i)}\boldsymbol{x})$からカテゴリ$i$ における新規文書$\boldsymbol{x}$ にカテゴリi における新規文書$\boldsymbol{x}$ に対する理想符号長$L_i(\boldsymbol{x})$ を次式で定義する.

$$L_i(\boldsymbol{x}) = L(\tilde{x}^{(i)}\boldsymbol{x}) - L(\tilde{x}^{(i)}). \quad (2)$$

最終的に新規文書 $\boldsymbol{x}$ が所属するカテゴリの推定値は, 次式により与えられる.

$$\hat{c} = \arg\min_i L_i(\boldsymbol{x}). \quad (3)$$

すなわち, 最も理想符号長を小さくするカテゴリに分類を行う. さて式(1), (2)より

$$L_i(\boldsymbol{x}) = L(\tilde{x}^{(i)}\boldsymbol{x}) - L(\tilde{x}^{(i)}) = -\log_2 \frac{P_w^{\lambda}(\tilde{x}^{(i)}\boldsymbol{x})}{P_w^{\lambda}(\tilde{x}^{(i)})} = -\log_2 P_w^{\lambda}(\boldsymbol{x}|\tilde{x}^{(i)}), \quad (4)$$

が成り立つ. 従って式(3)による分類法は, カテゴリ内に所属する各文書は同一の確率モデルから生起することを仮定した下で, 新規文書の条件付き事後確率を最大とするカテゴリに分類している.

本手法において, 各カテゴリに対する学習データのファイル容量を同程度にすると, 正分類率がまんべんなく向上することが知られている[7].

3. サブカテゴリに対する半教師付き学習法による文書分類

カテゴリ$i \in \{1,2,\cdots,C\}$の学習文書の集合をX_iと書く. 前節の手法は, 各カテゴリが同一の確率モデルから生起することを仮定している. 一方, 実際の文書分類では, 1 カテゴリ中に複数のトピックが含まれていることがよくある. そこで, 学習文書の各カテゴリの文書集合X_iの中から, 少数の文書に対して手動でサブカテゴリを構築する. ここでカテゴリ$i \in \{1,2,\cdots,C\}$に対するサブカテゴリ数をk_iと書き, カテゴリiのサブカテゴリ$j \in \{1,2,\cdots,k_i\}$ に含まれる学習文書の集合を$S_{i,j} = \{\tilde{x}_1^{(i,j)}, \tilde{x}_2^{(i,j)}, \cdots, \tilde{x}_{M_{i,j}}^{(i,j)}\}$ と記述する. ただし$S_{i,j}$は仮定より少数の文書しか含まないため, 各カテゴリに対して全学習文書をサブカテゴリに分類したい. そこで半教師付き学習によりこれを行うアルゴリズムを

A Study on Text Classification Using Context-Tree Weighting Algorithm for Semi-Supervised Leaning
Tomohiro OBATA†, Manabu KOBAYASHI† and Shigeyoshi WATANABE†
†Graduate School of Engineering, Shonan Institute of Technology, ‡Shonan Institute of Technology

以下に示す. ただし各サブカテゴリで連結した学習文書を$\tilde{X}_{i,j}$と表記し, 初期値は$S_{i,j}$中の全文書を連結して$\tilde{X}_{i,j} = \tilde{x}_1^{(i,j)}\tilde{x}_2^{(i,j)} \cdots \tilde{x}_{M_{i,j}}^{(i,j)}$と置く. また$X_i$ の中ですでにサブカテゴリを定め, $\tilde{X}_{i,j}$に含まれる文書の集合をY_i と表記する. また一度にカテゴリを判定する半教師文書の数をB で表す.

具体的には連結文書 $\tilde{X}_{i,j}$ のファイル容量が短いカテゴリに優先的に半教師文書を結合するものとする.

[サブカテゴリに対する半教師付き学習法]

(Step 1) $\tilde{X}_{i,j} = \tilde{x}_1^{(i,j)}\tilde{x}_2^{(i,j)} \cdots \tilde{x}_{M_{i,j}}^{(i,j)}, Y_i := S_{i,j}$.

(Step 2) 各カテゴリのサブカテゴリ未決定の学習文書集合$X_i \setminus Y_i$ からランダムに B 個の文書を選択する. またこのB 個の文書集合を Z_i とする.

(Step 3) すべての $i \in \{1,2,\cdots,C\}$ に対して以下を計算する.

各 $z \in Z_i$ に対し, 各カテゴリiのサブカテゴリ $j \in \{1,2,\cdots,k_i\}$ ついて$L_{i,j}(z|\tilde{X}_{i,j}) := L(\tilde{X}_{i,j}z) - L(\tilde{X}_{i,j})$ を計算する. また

$$\hat{j}_i(z) := \arg\min_j L_{i,j}(z) \quad (5)$$

を求める.

(Step 4) 以下を順に実行する.

(i) $(\tilde{\imath},\tilde{\jmath}) := \arg\min_{(i,j)} |\tilde{X}_{i,j}|$ とする.

(ii) $\tilde{\jmath} = \hat{j}_{\tilde{\imath}}(z)$ を満たす $z \in Z_{\tilde{\imath}}$ が存在しなければ (Step 5) へ. 存在するならばランダムに 1 つ選択し, $\tilde{X}_{\tilde{\imath},\tilde{\jmath}} := \tilde{X}_{\tilde{\imath},\tilde{\jmath}}z, Z_{\tilde{\imath}} := Z_{\tilde{\imath}} \setminus \{z\}, Y_{\tilde{\imath}} := Y_{\tilde{\imath}} \cup \{z\}$ として(i)へ.

(Step 5) (Step 4)(i)の $(\tilde{\imath},\tilde{\jmath})$ に対して $\tilde{\jmath} = \hat{j}_{\tilde{\imath}}(z)$ を満たす $z \in X_{\tilde{\imath}} \setminus Y_{\tilde{\imath}}$ が無ければ終了. そうでなければ(Step 2)へ. □

4. 計算機実験による評価

本節では毎日新聞の記事データ[9]を用いて, 2 節及び 3 節の分類手法に対する評価を行う. 実験では新聞のカテゴリとして「経済」,「国際」,「スポーツ」,「社会」の 4 カテゴリを対象とし, 各カテゴリに対して学習文書と新規テスト文書をそれぞれ用意した. 評価のための新規テスト文書数は, 各カテゴリに対して 400 ずつとする. また全カテゴリの合計の学習文書数$N = \sum_{i=1}^{C} N_i$を 400（カテゴリによって文書数が異なる点に注意）とする.

次に, サブカテゴリを構築する 3 節の手法の結果を示す. ここで各カテゴリのサブカテゴリ数を$j_i = 2$とし, 手動でサブカテゴリに分類を行った初期の学習文書の総数$M = \sum_{i,j} M_{i,j}$をそれぞれ 40, 200, 400（サブカテゴリによって文書数が異なる点に注意）とした. ただし簡単のため一度にカテゴリを半教師学習する文書数BをMと設定した. なおサブカテゴリとして「経済」を「政治的な経済記事」と「企業の経済記事」.「国際」を「政治」と「その他の国際記事」,「スポーツ」を「文化的な記事」と「試合結果」,「社会」を「事件, 事故」と「その他の社会記事」に手動で分けサブカテゴリを構築した. 表 1 にそれぞれの平均正分類率を示す. $M = 400$の時, すべての学習文書を手動でサブカテゴリに分類していることを表す.

表 1:サブカテゴリを用いた文書分類の平均正分類率(%)

M	従来	40	200	400
	83.3	70.6	80.8	81.1

表 1 を見るとサブカテゴリに分けた効果がほとんど見られず, 有効性が確認できない. これはカテゴリ「経済」と「国際」の中にそれぞれ「政治」に関するサブカテゴリを作成しており, これらの分類がうまくいっていないことが影響している.

5.まとめと今後の課題

本研究では, 各カテゴリ中に複数の確率モデルが混在する場合を想定し, サブカテゴリを構築するサブカテゴリに対する半教師付き学習法を提案した. 新聞記事データを用いた計算機実験では残念ながら有効性を確認することができなかった. 今後はサブカテゴリの構築方法や文書長をうまく選択することにより, より正分類率を高める方法を検討する予定である.

参考文献

[1] C.M.ビショップ, パターン認識と機械学習, Springer, 2008.

[2] J.Ziv and N.Merhav, "A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification," IEEE Trans. On Information Theory, vor.39, no.4, pp.1270-1279, July 1993.

[3] F.M.J.Willems,Y.M.Shtarkov and T.J.Tjalkens,"The Conectext Tree Weighting Method: Basic Properties,"IEEETrans. On Information Theory, vol.41, no.3, pp.653-664,May 1995.

[4] F.M.J.Willems, "The Conectext Tree Weighting Method:Extensions,"IEEE Trans. On Information Theory, vol.44, no.2, pp.653-664,pp.792-798, May 1998.

[5] K.Sadakane, T.Okazaki and H.Imai, "Implementing The Context Tree Weighting Method for Text Compression,"Proc. of the IEEE Data Compression Conference, pp.123-132, 2000.

[6] Z. Dawy, J. Hagenauer and A. Hoffmann, "Implementing the context tree weighting method for context recognition, "Proc. of the IEEE Data Compression Conference, p.536, March 2004.

[7] 小畑智広, 池上裕之, 小林 学, 坂下善彦, "文脈木重み付け法による確率モデルを限定したテキストの自動分類", 電子情報通信学会論文誌(D), Vol.J95-D,No.10, Oct. 2012.

[8] 小畑智広, 小林 学, 坂下善彦, "文脈木重み付け法を用いた半教師付き学習による文書分類", 電子情報通信学会非線形問題研究信学技報, vol. 112, no. 389, NLP2012-112, pp. 49-53, Jan 2013.

[9] CD-毎日新聞94'データ集, 日外アソシエーツ, 1995.