2G2-3

文学作品推薦のための文書分類

Text Classification for Literature Recommendation

藤井遼[*1] 櫻井彰人[*1]
Ryo Fujii Akito Sakurai

[*1]慶應義塾大学大学院理工学研究科
Graduate School of Science and Engineering, Keio University

In text classification for literature recommendation, there are several problems. First, zero-frequency problems occur more frequently than commonly observed in other text classification. We tried "m-estimation" known to work well for smoothing but found that it worked as well as conventional Laplace correction. Second, important features of literature liked (or hated) are varied among users. Feature selection is inevitable since important features vary among users and are buried in common features. We found that "Bi-Normal Separtion" is the best. Third, importance of literatures must be weighted to reflect relative importance. We evaluated these methods and parameters on data obtained from six subjects on Aozora-Bunko.

1. はじめに

今日の機械学習手法の発展により、商品推薦サービスが広まっている。しかしその多くは商品の属性を用いたものであり、より細やかなユーザーの好みには必ずしも対応できているとは言えない。ここではユーザーの選好情報に基づき特徴選択を行い、選択した特長を用いて文学作品の分類器を作成し、推薦に使用することを考える [Mooney 00]。

本課題においてはゼロ頻度問題が他の文書分類問題より頻繁に発生するため、これに対する対処が精度に大きく影響する。本論ではスムージングとして Laplace correction と m-estimation [*1][Cestnik 91] という手法を比較し、Laplace correction が本課題に適していることを示す。

また、人間が文学作品に好悪の評価を与える際、何によって評価を下しているかは人によって異なると考えられる。よって、ユーザーの選好によって文書を分類するには、使用する特徴を事前に選択できない。したがって、分類に使用する特徴を学習する必要がある。本論では特徴選択手法を比較し、BNS（Bi-Normal Separation）が最も優れていることを示す [Yang 97][Forman 03]。

さらに、文書に対する好悪は一様ではなく程度差がある。本論では単語の確率の推定へのその重みを導入を試みるが、精度には寄与しないことを示す。

最後に、選好に基づく分類に適した手法を組み合わせることで、高い精度で分類が出来ることを示す。

2. Naive Bayes 分類器

文書分類として bag of words モデルを使用する。Bag of words モデルに対する有効な分類アルゴリズムとしては Naive Bayes 分類器がある。Naive Bayes 分類器は精度は必ずしも高くないもののその高速性で他の高精度のアルゴリズムに対して優位に立っている。また再学習が容易なため、読書履歴によって学習データが随時変化する中で、大量の書籍を分類して推薦図書を探すという本論の目的に適している。

連絡先: 藤井 遼, 慶応義塾大学理工学研究科, 横浜市港北区日吉 3-14-1, 045-563-1141, roy@ae.keio.ac.jp

*1 ロバスト推定量 M-estimator を得る方法として知られている M-estimation とは異なる。

分類先クラス $c \in \{1,2\}$ はそれぞれ好き、嫌いな文書クラスを表すとする。学習データが文書の集合 $B = \{b_j | i \leq j \leq n\}$ で与えられ、文書 b_j の所属クラス $c_j \in \{1,2\}$ が既知であるとする。未知文書 b_d に対して、その文書が分類されるクラスは以下で与えられる。

$$c(b_d) = \arg\max_c [\log p(c) + \sum_{i=1}^{l} f_{id} \log p_c(i)] \quad (1)$$

ここで、$1 \leq i \leq l$ は学習データ内の語彙を表す。また、$p(c)$ はクラス c の文書が得られる確率、f_{id} は文書 b_d 中の単語 i の出現回数、$p_c(i)$ は単語 i がクラス c において出現する確率である。

3. スムージング

確率 $p(c)$, $p_c(i)$ の事前分布を仮定しない最尤推定値は、

$$\hat{p}(c) = \frac{|\{j|c_j = c\}|}{n} \quad (2)$$

$$\hat{p_c}(i) = \frac{a_{ci}}{N_c} \quad (3)$$

$$a_{ci} = \sum_{j=1}^{n} \{(1 - |c_j - c|) f_{ij}\},\ N_c = \sum_{i=1}^{l} a_{ci} \quad (4)$$

となる。ここで a_{ci} はクラス c における単語 i の出現度数、N_c はクラス c の単語度数である。

しかし、学習データにおいて片方のクラスにしか出現しなかった特徴については、このままでは $\hat{p_c}(i) = 0$ となってしまう。これはゼロ頻度問題と呼ばれ、スムージングと呼ばれる手法で最尤推定値を補正する。これは何らかの事前分布を仮定することに相当する。代表的なスムージング手法に以下のものがある。[Cestnik 91]

- Laplace correction

$$p_c(i) = \frac{a_{ci} + 1}{N_c + l} \quad (5)$$

- m-estimation

$$p_c(i) = \frac{a_{ci} + p_i m}{N_c + m},\ p_i = \frac{1}{l} \tag{6}$$

ここで m は m-estimation におけるパラメータである。

4. 特徴選択

分類において全ての単語を特徴として用いるのではなく、特徴的な単語だけを用いる方が精度がよい。一方、文学作品に対する好き嫌いに、文学作品中に表れる単語や言い回しも大きく影響している可能性がある。ユーザーによって、どのような単語が選好を決定する重要な情報なのかは異なるため、その重要度は学習データ中での単語の分布から決定すべきだと考えられる。

文書分類に使われる特徴選択のための指標には次のようなものがある [Forman 03]。

対象とする単語が出現したクラス c の文書数を o_c で、同様に出現しなかった文書数を q_c、クラス c の文書数合計を n_c で表す。

- Document frequency (DF) :

$$o_1 + o_2 \tag{7}$$

- 情報量利得 (IG) :

$$e(n_1, n_2) - \frac{o_1 + o_2}{n} e(o_1, o_2) - \frac{q_1 + q_2}{n} e(q_1, q_2) \tag{8}$$

ただし、$e(x, y) = -\frac{x}{x+y} \log \frac{x}{x+y} - \frac{y}{x+y} \log \frac{y}{x+y}$

- カイ 2 乗統計量 (χ^2)

$$\begin{array}{c} t(o_1, (o_1 + o_2)\frac{o_1+q_2}{n}) + t(o_2, (o_1 + o_2)\frac{o_2+q_1}{n}) \\ + t(q_1, (q_1 + q_2)\frac{o_2+q_1}{n}) + t(q_2, (q_1 + q_2)\frac{o_1+q_2}{n}) \end{array} \tag{9}$$

ただし、$t(x, y) = \frac{(x-y)^2}{y}$

- Bi-Normal Separation (BNS):

$$|F^{-1}(o_1) - F^{-1}(o_2)| \tag{10}$$

ただし、$F^{-1}(x)$ は標準正規分布の累積密度逆関数

これらの値がパラメータとして与えられた閾値を上回る単語だけを分類に使用する。この中で、DF は学習データの「各文書がどのクラスに所属しているか」という情報を特徴選択に使用しない。一方、その他の手法ではその情報を用いる。

5. 重み付け

文書に対する選好は単純な好き嫌いの 2 値ではなく、$[0, 1]$ の範囲の値をとると考えるのが実際に近いと考えられる。そこで、文書に対する選好度の重み w_j を考える。重みを推定に反映させるために、a_{ci} および N_c を次のように変更する。

$$a_{ci} = \sum_{j=1}^{n} \{(1 - |c_j - c|) w_j f_{ij}\} \tag{11}$$

$$N_c = \sum_{i=1}^{n} a_{ci} \tag{12}$$

このようにすると $p_c(i)$ は依然として確率である一方、文書間の重みが確率の推定に反映される。この後、スムージングは 3 節のように行う。w_j の推定方法については、6 節に記す。

6. 実験

選好に基づいた実験データを得るために、文学作品として青空文庫 [aozora] のデータを使用した。

6 名に対して文書を自由に選択してもらい、読後の感想を「大好き・好き・嫌い・大嫌い」の四段階で評価するアンケートを行った。その結果、総書籍数（正味）が 184、トークン種類が 62482、6 組のデータを得た。その結果を表 1 に示す。行は人に、列は評価値に対応する。表の各要素は各人（行）が評価値（列) を与えた文学作品の数である。

	大好き	好き	嫌い	大嫌い	合計
A	5	17	12	11	45
B	11	12	10	11	44
C	10	10	12	15	47
D	10	10	16	10	46
E	10	13	12	15	50
F	7	14	13	16	50

表 1: 選好データ概要

「大好き」「大嫌い」についてはパラメータとして重み $w_j = w$ を導入し、「好き」「嫌い」については $w_j = 1$ とした。

それぞれに対してスムージング・特徴選択・パラメータ w の異なる文書分類器を作成し、one versus others で各文書について所属クラスを推定し、正答率

$$\frac{|\{j | c(b_j) = c_j\}|}{n} \tag{13}$$

を求めた。特徴選択の閾値は、全学習データでの最大値と最小値の間を 20 分割して順にとり、それぞれの指標で最も正答率の良いものを用いた。閾値は、DF では 4、IG では 0.13、χ^2 では 0.6、BNS では 0.2 となった。m-estimation のパラメータ m は $m = \{0.01, 0.1, 1, 2, 4\}$ で比較し、平均的に正答率の高かった $m = 0.01$ を採用した。

以上の条件で、6 人の正答率の平均値でその結果を比較した。行はスムージング法に、列は特徴選択指標（閾値）に対応する。表の各要素はスムージング法と特徴選択の組み合わせでの正答率の平均値（%）である。

実験結果を表 2、表 3 に示す。

	DF	IG	χ^2	BNS
Laplace corrrection	64	67	68	71
m-estimation	64	61	62	62

表 2: $w = 1$(重み無し) の正答率

	DF	IG	χ^2	BNS
Laplace correction	64	67	64	66
m-estimation	62	59	61	61

表 3: $w = 2$(重み有り) の正答率

7. 考察

スムージング法として、m-estimation を使用するより Laplace correction を使用する方が精度が良いことが分かった。これは、m-estimation では推薦問題においてはスムーズ不足であるためだと考えられる。

特徴選択手法は BNS が最も良い結果であった。BNS は正規分布を仮定したときの平均値の差を表現するものであり、これが「単語が分類に関係あるか否か」をうまく表していたと考えられる。

文書ごとの選好度に基づく重みを加えても精度はよくならず、むしろ悪くなることが分かった。この方法では文書の選好度をうまく反映できていないのだと考えられる。

8. 今後の課題

現在はまだ各単語を同等に扱っているため、たとえ特徴選択を行ったとしても、どの単語がどの程度選好に影響しているかという情報を扱えていない。したがって、Naive Bayes 以外の他のモデルにも今回の知見を生かすことで精度の向上を目指す必要がある。

参考文献

[Mooney 00] Raymond J. Mooney, Loriene Roy. Content-based book recommending using learning text categorization. In: Proceedings of the fifth ACM conference on Digital libraries, ACM, pp. 195-204, 2000

[Cestnik 91] Bojan Cestinik, Ivan Bratko. On Estimating Probablities in Tree Pruning. In: Proceeding of European Working Sessions on Learning EWSL 91, Lecture Notes in Artificial Intelligence, vol. 482. Springer, Berlin, pp. 138-150, 1991

[Yang 97] Yiming Yang, Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In: Proceeding of the Fourteenth International Converence on Machine Learning. pp. 412-420, 1997

[Forman 03] George Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research, vol. 3, pp. 1289-1305, 2003

[aozora] 青空文庫：http://www.aozora.gr.jp/