文書分類における補集合を併用した Naive Bayes

伊藤 裕佑(東京農工大学工学部)†
古宮 嘉那子(東京農工大学工学研究院)
小谷 善行(東京農工大学工学研究院)

Naive Bayes using The Complement Set in Text Classification

Yusuke Ito (Department of Computer and Information Sciences Faculty of Engineering),
Kanako Komiya (Institute of Engineering Tokyo University of Agriculture and Technology),
Yoshiyuki Kotani (Institute of Engineering Tokyo University of Agriculture and Technology)

１．はじめに

これまで文書分類に関する研究は数多くなされてきており，これらの研究において Bayes のアプローチがよく用いられている．Naive Bayes を発展させた研究として，Rennie らによる「補集合」を用いた Complement Naive Bayes [1]や、Komiya らの Negation Naive Bayes[2]がある．本研究では，Naive Bayes と Negation Naive Bayes に注目し，補集合を利用した新しい手法を提案する．それには，Naive Bayes と Negation Naive Bayes の統合およびクラスごとの学習量による選択を行う．

２．関連研究

Andrew は Naive Bayes を適用して分類を行う際の事象モデルとして，多項モデルと多変量ベルヌーイモデルの違いを述べ，分類結果から多項モデルの優位性を示している[3]. Komiya らは、Rennie らによる Complement Naive Bayes という手法に注目し，Negation Naive Bayes を提案している．本研究では多項モデルの Naive Bayes と Negation Naive Bayes に注目し，新しい手法を開発する．

３．Bayes の定理を用いた既存の文書分類法

本研究で提案する手法の基礎となる分類器について述べる．これらの分類器は本研究で文書分類の実験を行う際の比較対象となる．

分類器が分類先を推定する際には，Bayes の定理を利用した推定式から事後確率 $P(c|d)$ が最大となるクラス $\hat{c}$ を求める．Naive Bayes は $P(c|d)$ を最大化するように Bayes の定理をそのまま適用した分類器である．これはクラスごとに与えられた文書をそのまま学習に利用する． Negation Naive Bayes は「クラスに属さない文書」，つまり「補集合」を考えることでクラスごとの学習事例数を増やすように工夫している．

３．１　Naive Bayes

確率モデルによる文書分類において，分類対象となる文書を d，ある一つのクラスを c とするとき，事後確率 $P(c|d)$ を最大化するクラス $\hat{c}$ を求めればよい．

Naive Bayes では，$P(c|d)$ に Bayes の定理を適用するが，文書の取り出される確率 $P(d)$ がすべてのクラスについて一定であるので，Naive Bayes はクラスの出現確率 $P(c)$ と各クラスにおける文書の出現確率 $P(d|c)$ の積を最大化するクラスを推定する．ただし，文書 d は $(w_1, w_2, ..., w_n)$ のような単語列からなる．

$$\hat{c} = \arg\max_c P(c) \prod_{i=i}^{n} P(w_i|c) \quad (1)$$

† 50010268502@st.tuat.ac.jp

３．２　Negation Naive Bayes

Naive Bayes による文書分類では，ある一つのクラス c に関する学習に「c に属する訓練事例」を用いていた．それを発展させた分類器として Complement Naive Bayes があり，「c に属さない訓練事例」すなわち「$\bar{c}$ に属する訓練事例(補集合)」を用いて学習する．しかし，Rennie らの研究ではアプローチが特徴的であるものの数学的根拠がない[1]．そのため，Komiya ら[2]は補集合を用いたモデルに基づく方法を開発してきている．

事後確率 $P(c \mid d)$ を最大化するクラス $\hat{c}$ を求める式を変形する．

$$\hat{c} = \arg\max_c P(c \mid d) = \arg\max_c (1 - P(\bar{c} \mid d)) = \arg\min_c P(\bar{c} \mid d) \tag{2}$$

次に，Bayes の定理を用いて Naive Bayes と同様に変形する．

$$\hat{c} = \arg\min_c P(\bar{c}) P(d \mid \bar{c}) \tag{3}$$

したがって，文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c} = \arg\min_c P(\bar{c}) \prod_{i=1}^{n} P(w_i \mid \bar{c}) \tag{4}$$

４．補集合を併用した Naive Bayes の提案

本研究では，Negation Naive Bayes と同様に Bayes の式を変形することで新しい方法を考え，ここではそれを「Universal-set Naive Bayes」と呼ぶ．また，各クラスの $P(c)$ によって Naive Bayes と Negation Naive Bayes を選択して処理する方法を新たに考え, それを「Selective Naive Bayes」と呼ぶ.

４．１　Universal-set Naive Bayes

事後確率 $P(c \mid d)$ の推定に $P(d)$ は不要であるとして式(1), (2)は導出されている．しかし，ここでは式(6)を $P(d)$ について解くことで新たな分類器を考える．式(6)は $P(c \mid d)$ と $P(\bar{c} \mid d)$ を足し合わせた全体の確率

$$P(\text{全体} \mid d) = P(c \mid d) + P(\bar{c} \mid d) = 1 \tag{5}$$

に Bayes の定理を適用している．

$$\frac{P(c)P(d \mid c)}{P(d)} + \frac{P(\bar{c} \mid d)}{P(d)} = 1 \tag{6}$$

式(6)を変形することで $P(d) = P(c)P(d \mid c) + P(\bar{c})P(d \mid \bar{c})$ が得られ, Bayes の式を次のように書きかえられる．

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} = \frac{P(c)P(d \mid c)}{P(c)P(d \mid c) + P(\bar{c})P(d \mid \bar{c})} = \frac{1}{1 + \dfrac{1}{\dfrac{P(c)P(d \mid c)}{P(\bar{c})P(d \mid \bar{c})}}} \tag{7}$$

式(7)の右辺の分数を突き詰めていくと，左辺 $P(c \mid d)$ の最大化が右辺 $\dfrac{P(c)P(d \mid c)}{P(\bar{c})P(d \mid \bar{c})}$ の最大化となるので，Universal-set Naive Bayes は文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c} = \arg\max_c \frac{P(c)}{P(\bar{c})} \prod_{i=1}^{n} \frac{P(w_i \mid c)}{P(w_i \mid \bar{c})} \tag{8}$$

４．２　Selective Naive Bayes

Negation Naive Bayes はばらつきを抑えて Naive Bayes より分類性能を向上させているが，補集合を学習に用いることで逆に学習事例数を減らしてしまう場合がある(図 1)．これを $P(c)$ に基づいてクラスごとに式(1)，(2)で選択し，文書分類を行なう．このとき，0.5 をしきい値とすることで，学習する事例数がより大きくなるように選択する．

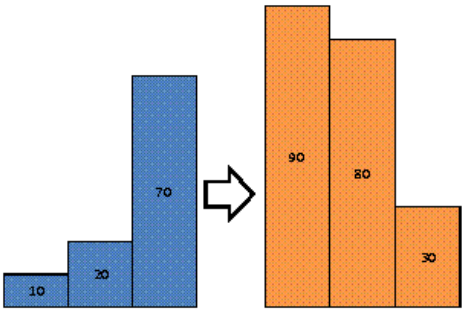

図 1　補集合をとる前後の 3 クラスの学習事例数(例)

分類先のクラスを推定する際には，Naive Bayes や Negation Naive Bayes 単独のときと異なり，$P(d)$ を求める必要がある．この $P(d)$ は以下に示す式(9)のように異なる導出がある．これらは式(6)を変形していくことで得られる．

$$P(d)=\sum_{c}P(c)\prod_{i=1}^{n}P(w_i\mid c)=\frac{1}{|C|-1}\sum_{c}P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c}) \tag{9}$$

したがって，文書 d の属するクラス $\hat{c}$ を次の式で推定する．

$$\hat{c}=\arg\max_{c}\begin{cases}\dfrac{P(c)\prod_{i=1}^{n}P(w_i\mid c)}{\sum_{c}P(c)\prod_{i=1}^{n}P(w_i\mid c)}, & P(c)\text{が}0.5\text{以上}\\ 1-(|C|-1)\dfrac{P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c})}{\sum_{c}P(\bar{c})\prod_{i=1}^{n}P(w_i\mid \bar{c})}, & \text{その他}\end{cases} \tag{10}$$

５．実験

ここでは，提案手法の性能を評価するための実験を行い，その結果を示す．

比較する既存手法として，ベースラインの Naive Bayes，および Negation Naive Bayes[2] を取り上げる．なお，ここでは Naive Bayes，Negation Naive Bayes，Universal-set Naive Bayes，Selective Naive Bayes をそれぞれ NB，NNB，UNB，SNB と略記する．

分類性能を評価する実験に用いるコーパスには「現代日本語書き言葉均衡コーパス(BCCWJ)」の一部を用いる(図 2)．五つのジャンルがあり，Yahoo!知恵袋，白書，書籍，雑誌，新聞である．BCCWJ の実験は bag-of-words に加工済みのデータを用いて文書分類を行なっている．5 分割交差検定で実験を行うが，データセットについては 5 クラスの場合だけでなく，3 クラスのデータセットを作った場合も実験を行った．5 クラスよりもシンプルな 3 クラスの分類実験を行なって提案手法と既存手法の違いを確認する．

５．１　実験方法と 3 クラス分類のデータセット

実験は各データセットについて 5 分割交差検定を行い評価する．5 クラスのデータすべてを使った 5 クラス分類と，よりシンプルな 3 クラス分類 10 通りをそれぞれ実験する．5 ジ

ャンルからクラスを 3 つ選ぶ組合せが 10 通りとなる.

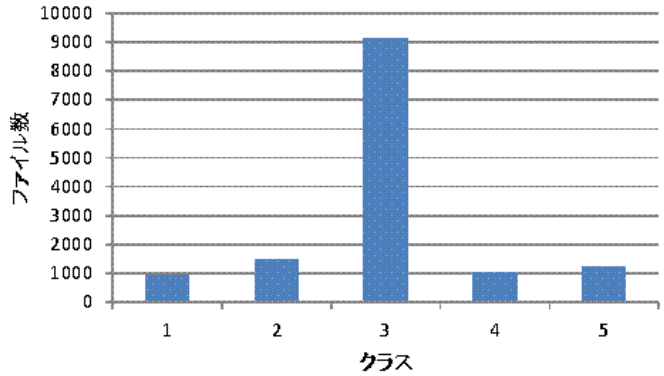

図 2　クラスごとの文書ファイル数(1: Yahoo!知恵袋，2:白書，3:書籍，4:雑誌，5:新聞)

５.２　実験結果

評価実験の結果は表 1 および 2 のとおりであり，シンプルな 3 クラス分類について提案手法が優れており，その違いが有意であることが確かめられた．また，どちらの表からも提案手法の一つである UNB が優れていることがわかり，適用できる問題が広範囲である可能性を持つ.

表 1 5 クラス分類における各手法の評価指標

	NB	NNB	UNB	SNB
Precision	0.7036	0.5574	**0.7442**	0.2505
Recall	**0.8314**	0.3985	0.7573	0.2877
F-measure	**0.7486**	0.3356	0.7381	0.2638
Accuracy	0.7798	0.6683	**0.8048**	0.4500

表 2 3 クラス分類における各手法の評価指標の平均値

	NB	NNB	UNB	SNB
Precision	0.6518	0.6230	**0.6687**	0.5455
Recall	**0.7078**	0.5755	0.7034	0.4748
F-measure	0.5425	0.4744	**0.5641**	0.4550
Accuracy	0.5577	0.5587	**0.5933**	0.5711

６．考察・今後の課題

本研究の開発した新しい手法が文書分類において既存手法に比べて有効であることが示された．Universal-set Naive Bayes が良い手法であることは確かめられたが，Selective Naive Bayes は今後，他の手法との違いをデータセットを変えて詳しく確かめる必要がある.

文　献

J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger(2003) Tackling the Poor Assumptions of Naive Bayes Text Classification, ICML2003, pp.616-623

Kanako Komiya, Naoto Sato, Koji Fujimoto and Yoshiyuki Kotani, (2011)Negation Naive Bayes for Categorization of Product Pages on the Web, 2011 (RANLP 2011), pp586-591

Andrew McCallum, Kamal Nigam(1998) A Comparison of Event Models for Naive Bayes Text Classification, AAAI/ICML-98 Workshop on Learning for Text Categorization, pp.41-48