社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－166 (14)
2005／3／11

単語間の依存性を考慮した ナイーブベイズ法によるテキスト分類

花井拓也 山村毅

愛知県立大学 〒480-1198 愛知県愛知郡長久手町大字熊張字茨ヶ廻間 1522-3

本稿では，ナイーブベイズ法によるテキスト分類において，2単語間の依存性を考慮することでより正確な確率を計算する手法について述べる．ナイーブベイズ法では，クラスが与えられたとき各要素間に独立性を仮定しているが，テキスト分類においてはカテゴリが与えられたとき各単語の生起の間には完全な独立性は成り立たない．そこで，ナイーブベイズ法による確率計算において，依存性があると考えられる単語の組については，個々の単語の生起確率の積のかわりに同時確率を用いて計算を行う手法を提案し，Reuters-21578 コーパスを用いた分類実験によりその有効性を評価する．

Text categorization using Naive Bayes method that considers the dependency between words

Takuya Hanai and Tsuyoshi Yamamura

Aichi Prefectural University. 1522-3 Ibaragabasama, Kumabari, Nagakute-cho, Aichi-gun, Aichi, 480-1198, JAPAN

This paper describes the technique of calculating accurate probability in the text categorization using Naive Bayes method, taking the dependability between two words into consideration. In Naive Bayes method, independency is assumed between each element, but that independency isn't fully saticefied in the text categorization. Therefore we calculate accurate probability for dependable word pairs. We evaluate the validity of our method by the categorization experiment using Reuters-21578 corpus.

1 はじめに

インターネットの普及により膨大な情報が流通している現在，大量の情報の中から目的とする情報を効率的に取得する方法が求められており，そのための一つの手段としてテキスト分類技術が重要視されている．これは，情報をあらかじめ定められたカテゴリに分類し整理しておけば，その中から目的とする情報を探し出すことは比較的容易になるからである．

テキスト分類技術は，たとえばインターネットにおいて Web ページをあらかじめ階層的に分類しているディレクトリ型検索エンジンや，送られてくる電子メールを「ふつうのメール」と「スパム」に分類することで，不特定多数に一方的に送りつけられる広告や勧誘等の電子メールをシャットアウトするスパムメールフィルタなど [5] で利用されている．また，キーワード等で指定することにより自分が興味をもつ分野の記事だけを配信してくれる電子メールニュースサービス，過去の購買履歴から自分の興味にあう本を推薦してくれるオンライン書店，顧客からの電子メールを専門分野別にオペレータへ振り分けるオンラインヘルプデスクなどにも，応用することができる．

テキスト分類技術の歴史的な変遷を振り返ると，1980 年代後半までは知識工学的アプローチ，すなわち人手で分類規則を書く方法が主流であった．しかし，1990 年代に入ると，大量のテキストデータが利用可能になったことや，コンピュータの性能が大幅に向上したことから，機械学習的アプローチ，すなわち人手によりカテゴリラベルを付加したテキストデータから自動的に分類器を作成する方法が，分類精度・省力製・保守性に優れているために，主流となった [4]．機械学習を行うための代表的なモデルとしては，確率モデル，規則に基づくモデル，ベクトル空間モデルなどがある．

確率モデルにもとづく方法の中で最もよく利用される方法にナイーブベイズ法がある．この方法では

クラスのもとで各属性値の間の独立性を仮定しているが，テキスト分類においては各単語の生起の間に完全な独立性は成り立たない．すべての単語間の依存性を考慮することは困難であるので，本稿では，依存性が強いと考えられる2単語の組についてのみ依存性を考慮し，より正確な確率を計算することにより分類精度の向上を考える．

ナイーブベイズ分類器の精度を向上させる研究として，文献 [6] では文書クラスタリングと単語クラスタリングを用いている．テキスト分類において，カテゴリは人間によって用意されたものであり，その背後に確率的構造が置かれているわけではない．同じカテゴリの文書は同一の分布から生成されるという仮定が誤りであるとし，文書に対してクラスタリングを用いることによりこれを解消しようとしている．また，単語クラスタリングでは確率モデルにおけるデータスパースネスの問題を軽減している．そして実験の結果，確率モデルの是正の効果による分類精度の向上を報告している．

以下に本稿の構成を示す．2章でナイーブベイズ法によるテキスト分類について述べ，3章で単語間の依存性を考慮したナイーブベイズ法による分類を提案する．4章で提案手法に基づいた分類システムを作成，分類実験を行い．5章で考察を行う．最後に6章でむすびとする．

2 ナイーブベイズ法によるテキスト分類

2.1 テキスト分類

テキスト分類 (text categorization) とは，テキストをその内容に従ってあらかじめ決められたカテゴリに分類することをいう．テキスト分類の基本的な手続きは以下のようになる．

1. 各文書を単語等に基づく多次元のベクトルで表現する．
2. 訓練データ (人手によりカテゴリラベルを付与された文書集合) を利用して，文書を表すベクトルとカテゴリの対応関係について教師あり学習を行う．
3. 未知文書 (分類を行う文書) にもっとも類似したカテゴリを付与する．

2.2 情報利得を用いた素性選択

文書のベクトル表現において，stop-word を除いたすべての単語を使用した場合，計算量の増加や訓練データのオーバーフィットといった問題が生じる．そこで通常，単語の中から特に分類にとって重要である単語 (これを素性と呼ぶ) を何らかの基準のもとで抽出する．この素性選択に用いられる様々な評価基準として，文書頻度，相互情報量，情報利得，カイ2乗検定などがある [3]．本研究では素性選択法としてカテゴリを選ぶうえでの各単語の情報利得を用いた．

情報利得 (IG:Information Gain) は文書に単語が出現するか否かが与えられたときに，文書が属するカテゴリの曖昧性がどれだけ減少するかを表す指標である．これは単語の出現の有無が文書の属するカテゴリの推定に役に立つときに大きな値を取るということであるので，情報利得 IG の値の大きい単語はテキスト分類において文書を表す素性として適当であると考えられる．カテゴリ集合 $\{c_1, ..., c_m\}$ に対する単語 w の情報利得 IG は次式で定義される．

$$\begin{aligned} IG(w) = &-\sum_{k=1}^{m} P(c_k) log P(c_k) \\ &+P(w)\sum_{k=1}^{m} P(c_k|w) log P(c_k|w) \\ &+P(\bar{w})\sum_{k=1}^{m} P(c_k|\bar{w}) log P(c_k|\bar{w}) \quad (1) \end{aligned}$$

ここで，

- $P(w)$:単語 w が文書に現れる確率
- $P(\bar{w})$:単語 w が文書に現れない確率
- $P(c_k)$:文書がカテゴリ c_k に属する確率
- $P(c_k|w)$:単語 w が文書に現れたうえで文書がカテゴリ c_k に属する確率
- $P(c_k|\bar{w})$:単語 w が文書に現れなかったうえで文書がカテゴリ c_k に属する確率

である．

2.3 ナイーブベイズ法による分類 [4]

確率モデルに基づくテキスト分類では，一般に，文書ベクトルを $\boldsymbol{x} = (x_1, ..., x_n)$，文書のカテゴリを c とするとき，事後確率 $P(c|\boldsymbol{x})$ を最大化するカテゴリ c を求めれば，文書分類の誤りを最小化できると考える．実際には，事後確率にベイズの規則 (Bayes rule) を適用し，すべてのカテゴリについて $P(\boldsymbol{x})$ は一定であることを考慮すると，結局，カテゴリの出現確率 $P(c)$ とカテゴリ別の文書ベクトルの出現確率 $P(\boldsymbol{x}|c)$ の積を最大とするカテゴリを求

めればよい.

$$\begin{aligned}\hat{c} &= arg\max_c P(c|\boldsymbol{x}) \\ &= arg\max_c \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})} \\ &= arg\max_c P(\boldsymbol{x}|c)P(c)\end{aligned}$$

この式において，$P(c)$ はカテゴリが付与された文書におけるカテゴリの相対頻度から容易に推定できるが，文書ベクトルはさまざまな値を取りうるので $P(\boldsymbol{x}|c)$ を直接推定するのは非常にむずかしい．そこで，カテゴリ c が与えられたときの文書ベクトル $P(\boldsymbol{x})$ の分布において，文書ベクトルの各要素が互いに独立と仮定する．

$$\begin{aligned}P(\boldsymbol{x}|c) &= P(x_1, ..., x_n|c) \\ &\approx \prod_{i=1}^{n} P(x_i|c)\end{aligned}$$

$P(x_i|c)$ はカテゴリが付与された文書における素性 x_i の相対頻度から推定できる．したがって，文書の分類は次式を最大化するカテゴリ $\hat{c}$ を選べばよい．

$$\hat{c} = arg\max_c P(c)\prod_{i=1}^{n} P(x_i|c) \qquad (2)$$

式 (2) に基づく方法を，ナイーブベイズ分類器と呼ぶ．

ナイーブベイズ法は入力文書がカテゴリに所属する確率を与えるので，音声認識結果のようにデータが雑音を含む場合や，ラベルありデータとラベルなしデータを組み合わせて学習する場合などには，理論的な拡張が容易であるという利点を持つ [2].

3 単語間の依存性を考慮した分類

3.1 手法の概要

ナイーブベイズ法では，カテゴリ c のもとですべての単語の生起が互いに独立であるという仮定を置いている．そのため，式 (2) では，カテゴリ c のもとでの各単語の生起についての確率 $P(x_i|c)$ を掛け合わせることで $P(x_1, ..., x_n|c)$ を計算している．しかし，自然言語ではある単語が別の単語の生起に影響を与える場合が往々にしてある．すなわち，互いの生起に関して依存関係がある単語があるということである．そこで本研究では，2 つの単語の間の依存関係を考慮し，より正確な確率を計算する．

2 つの単語の間に依存関係があれば，カテゴリ c のもとでの 2 つの単語の個々の生起についての確率 $P(x_i|c) \times P(x_j|c)$ と同時確率 $P(x_i, x_j|c)$ とは異なった値になるはずである．そこで，$P(x_i|c) \times P(x_j|c)$ と $P(x_i, x_j|c)$ とが大きく異なる単語の組を依存関係がある単語の組と考え，$P(x_i|c) \times P(x_j|c)$ のかわりに $P(x_i, x_j|c)$ を用いて計算を行う．

3.2 依存性の尺度

$P(x_i|c) \times P(x_j|c)$ と $P(x_i, x_j|c)$ のいずれを用いるかの判断には，それらの違いの尺度，すなわち x_i, x_j の生起の依存関係の強さの尺度の定義が必要となるが，これに，個々の単語の生起についての確率の積 $P(x_i|c) \times P(x_j|c)$ と同時確率 $P(x_i, x_j|c)$ との比を用いる．カテゴリ c のもとでの x_i と x_j の依存度 D を，$P(x_i|c) \times P(x_j|c)$ と $P(x_i, x_j|c)$ のうち大きい方の確率を小さい方の確率で割って得られる値であるとする．この値が大きいものほど 2 つの確率の違いが大きく，2 単語の間の依存関係が強いと考える．また，最小値は 1 であり，そのときはカテゴリ c のもとで x_i と x_j に依存性はないといえる．

$$D(x_i, x_j|c) = \begin{cases} \frac{P(x_i,x_j|c)}{P(x_i|c)\times P(x_j|c)} & \text{if } P(x_i|c) \times P(x_j|c) \leq P(x_i, x_j|c) \\ \frac{P(x_i|c)\times P(x_j|c)}{P(x_i,x_j|c)} & \text{if } P(x_i|c) \times P(x_j|c) > P(x_i, x_j|c) \end{cases} \qquad (3)$$

ただし，
$P(x_i|c) \times P(x_j|c) = 0$ のときは，$D(x_i, x_j|c) = 1$,
$P(x_i|c) \times P(x_j|c) \neq 0, P(x_i, x_j|c) = 0$ のときは
$P(x_i, x_j|c) =$ "きわめて小さな値" として計算する．

3.3 依存性を考慮した分類

依存関係の強い単語の組に対しては，すべて同時確率を考えて確率を計算したいが，これを行うことは，計算コスト上，また統計的にも問題が生じる．そこで，2 つの単語間の依存性のみを考えることにする．しかし，この場合，ある単語が複数の単語と依存性があるときには，どの単語との組み合わせにするのかを決める必要がある．例えば，単語 x_1, x_2, x_3 が依存関係にある場合，$P(x_1, x_2, x_3|c)$ を $P(x_1, x_2|c) \times P(x_3|c)$,$P(x_1, x_3|c) \times P(x_2|c)$,$P(x_2, x_3|c) \times P(x_1|c)$ のいずれで近似するかを決める必要がある．本研究では，より依存関係の強い組を選択することにする．

3.4 依存性を考慮する組の絞込み

上述のように 2 つの単語の依存性のみを用いたとしても，素性選択した単語すべてから 2 単語の組み合わせを考え，各カテゴリごとに個々の単語の生起についての確率の積と同時確率の違いを計算し

ていくため，同時確率を用いる組の候補が非常にたくさんになってしまう．そこで，最初に2単語の組み合わせの分類における重要度を求め，それが一定値以上のものについてのみ同時確率を用いる組の候補とする．具体的には情報利得を用いて重要度を決める．

3.5 アルゴリズム

ここでは，提案手法を用いた分類の手順を説明する．基本的にはテキストのベクトル表現，各確率の学習，分類に分けることができる．以下に，それぞれのフェーズにおいての手順を説明する．

3.5.1 テキストのベクトル表現

ナイーブベイズ法で分類を行うために，stop-wordを除いたすべての単語の中からカテゴリの分類に寄与しそうな単語（素性）集合 $\boldsymbol{w} = \{w_1, ..., w_n\}$ を選択し，その単語の有無を要素として各文書をベクトル $\boldsymbol{x} = (x_1, ..., x_n)$ で表現する．すなわち，単語 $w_i, (i = 0, ..., n)$ が文書に出現すれば $x_i = 1$，しなければ $x_i = 0$ とする．

3.5.2 学習

学習では，訓練データを用いて，各カテゴリごとに以下を計算する．
$i, j = (1, ..., n), i \neq j$ として，

1. 集合 $N = \emptyset$ を用意．
2. $P(c)$ の計算．
3. $\boldsymbol{x}$ 中のすべての要素 $x_i (x_i = 1, x_i = 0$ の2通りがある）について $P(x_i|c)$ の計算．
4. $\boldsymbol{x}$ 中のすべての要素の組み合わせ (x_i, x_j) $((x_i, x_j) = (1,1), (1,0), (0,1), (0,0)$ の4通りがある）について，
   (a) $P(x_i, x_j|c)$ と $D(x_i, x_j|c)$ の計算．
   (b) $D(x_i, x_j|c) \geq$ 閾値α である組 (x_i, x_j) を集合 N に追加．

3.5.3 分類

分類は，以下の手順で行う．

1. 集合 $Bi = \emptyset, Uni = \emptyset$ を用意する．
2. 未知文書 $\boldsymbol{d}$ の (x_i, x_j) の部分と一致するかどうかを $D(x_i, x_j|c) \in N$ の大きい組から順に調べていく．
3. 一致する組 (x_i, x_j) が見つかったら，その組を Bi に追加し，N から単語 w_i と w_j を含む組を全て取り除く．
4. すべての $(x_i, x_j|c) \in N$ がチェックし終わるまでステップ1～2を繰り返す．
5. Bi に含まれる組で使われていない単語は Uni に追加する．
6. 下式を計算して，分類を行う．

$$\hat{c} = arg\max_{c} P(c) \prod_{(w_i, w_j) \in \boldsymbol{Bi}} P(x_i, x_j|c) \times \prod_{w_i \in \boldsymbol{Uni}, w_i \in \boldsymbol{d}} P(x_i = 1|c) \times \prod_{w_i \in \boldsymbol{Uni}, w_i \notin \boldsymbol{d}} P(x_i = 0|c) \tag{4}$$

4 実験と評価

4.1 実験に使用するコーパス

実験には Reuters-21578 を用いた．Reuters-21578 の訓練記事とテスト記事の振り分け方法は，"The Modified Apte(ModApte) Split"[1] に従った．また，本実験での設定を以下に示す．

- "TOPICS" を正解カテゴリラベルとする．
- "TOPICS" を持たない記事は使用しない．
- 本文を持たない記事は使用しない．
- 訓練記事集合で出現回数が5回以下の"TOPICS"は削除する．

その結果，本実験で使用する訓練記事は7036件，テスト記事は2725件となり，総トピック数は60となった．また，訓練記事，テスト記事の両方ともに対して，TreeTagger [*1]という品詞付与のできる形態素解析ツールを用いて複数形・語形変化への対応，数詞の除去を行った．また，SMART システムの stop-word リスト[*2]を用いて stop-word の除去を行った．

4.2 評価方法

精度の評価方法には，以下の正解率を用いた．

$$\text{正解率} = \frac{\text{正解した文書数}}{\text{全文書数}} \tag{5}$$

ここで，分類器が最も確率が高いとしたカテゴリが，正解カテゴリと一致した場合に正解とする．ま

[*1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html より入手可能
[*2] ftp://ftp.cs.cornell.edu/pub/smart/english.stop より入手可能

D の閾値	1	10	100	1000	10000	100000	1000000
考慮する組の平均数	344809	67907	57639	31001	6619	394	15
正解率 (%)	83.08	82.97	83.44	82.45	60.22	67.08	81.32

表 1: 提案手法による分類結果

IG の閾値	0.2	0.3	0.5	1	1.5	2	3
考慮する組の平均数	83871	43469	12223	3603	2102	1874	806
正解率 (%)	83.19	84.18	83.92	82.93	82.78	82.64	82.45

表 2: 考慮する組の絞込みを行った場合の分類結果

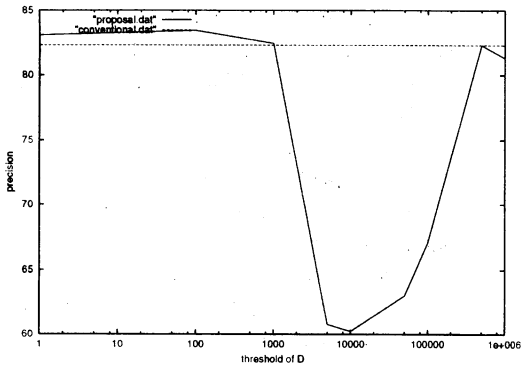

図 1: 提案手法による分類結果

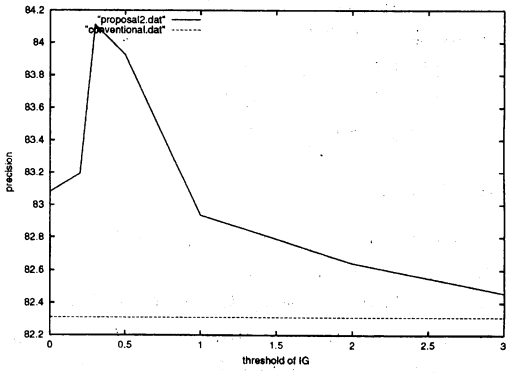

図 2: 考慮する組の絞込みを行った場合の分類結果

た，文書が複数の正解カテゴリを持つ場合は，そのうちのどれか 1 つに当てはまれば正解とする．

4.3 実験結果

通常のナイーブベイズによる分類で，素性数を 10 から 5000 まで変化させて正解率を調べたところ，素性数 700 のときが最も正解率が高く，82.31 %となった．そこで，提案手法においても素性数は 700 に設定して実験を行った．

式 (3) の各カテゴリのもとでの単語間の依存度 D の値が閾値 α 以上の組を用いるとして，α を 1 から 1000000 まで変化させて正解率を調べた結果を図 1 および表 1 に示す．図において実線は提案手法を，点線は通常のナイーブベイズによるものを示している．

正解率は閾値 α が 100 のときが最も高く，83.44 %の正解率を示した．これは通常のナイーブベイズより 1.13 %の向上である．また，α が 1000 以下では通常のナイーブベイズより正解率の向上が見られた．しかし，α がそれより大きいときには正解率は減少している．

次に，提案手法において考慮する組の絞込みを行った場合での分類を行った．$\alpha = 1$ とし，組み合わせの重要度の基準である情報利得の値が閾値 β 以上の組を用いるとして，β を変化させて正解率を調べた結果を図 2 および表 2 に示す．

正解率は $\beta = 0.3$ のときが最大で 84.18 %となった．これは通常のナイーブベイズと比較して 1.87 %の向上である．また，すべての β の値に対して正解率は通常のナイーブベイズより向上している．

5 考察

図 1 および表 1 より，提案手法では依存度の小さい組まで使った場合によい正解率が得られるが，依存度の大きい組のみを用いた場合には正解率が下がってしまうことがわかった．その原因としては，依存度の上位の組に $P(x_i|c) \times P(x_j|c)$ は 0 ではないが $P(x_i, x_j|c)$ は 0 であるという場合が多くあることが考えられる．そのような組が訓練データでは

出現しなかったがテスト文書では出現したという場合，文書がカテゴリ c に属する確率は一律にほぼ0になってしまうからである．

この他に，訓練データをよくモデル化し過ぎてしまっていることや，カテゴリは手作業で付与されたものであり，その背後に確率的構造が置かれているわけではないため，正確な確率計算が精度向上につながっていない等が考えられる．

また，図2および表2より，情報利得を用いて考慮する組を絞込んだ場合の方がこれを行わない場合に比べて正解率が高くなっている．このことから，2つの単語の同時生起確率を使ってナイーブベイズ分類の確率を変更する場合には，単に確率値の違いが大きい組をより正しく変更するよりも，分類において意味のある組の生起確率を正しく変更した方がよいということがわかった．

6 むすび

本研究では，ナイーブベイズ法によるテキスト分類において，単語の生起を完全には独立と仮定せず，2単語の間での依存関係を考慮して計算する方法を提案した．Reuters-21578 コーパスを用いて従来方式と提案手法の比較実験を行い，2単語間の依存関係を考慮することで正解率を向上させることができた．また，分類において重要な単語の組について依存関係を考慮することでより正解率を向上させることができた．

今後の課題としては次のものがある．まず，本稿では1つのデータセットのみを用いたが，他のデータセットを用いた実験も行う必要があると思われる．また，文献 [6] では，同一カテゴリ文書は同一の確率分布から生起したものだという仮定が誤りであると述べているが，正しい確率的構造においては，本研究の手法の効果がより顕著になる可能性がある．文書クラスタリングを用いて正しい確率的構造のもとでの実験も調査してみる必要があると思われる．

謝辞

本研究の一部は文部科学省科学研究費補助金 (課題番号 16200001) の支援による．

参考文献

[1] David D.Lewis, "Representation and Learning in Information Retrieval", Ph. D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA. 1992

[2] Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. , "Text classification from labeled and unlabeled documents using EM", Machine Learning, 39(2/3),103-134, 2000

[3] Yang, Y. and Jan O. Pederson, "A Comparative Study on Feature Selection in Text Categorization", Proc. 14th International Conference on Machine Learning, pp.412-420, 1997

[4] 金明哲，村上征勝，永田昌明，大津起夫，山西健司，"統計科学のフロンティア 10 言語と心理の統計"，岩波書店，2003

[5] 佐々木稔，新納浩幸，"文書分類を用いたスパムメール判定手法"，情報処理学会研究報告，NL-163, pp75-82, 2004

[6] 高村大也，松本裕治，"文書分類のための共クラスタリング"，情報処理学会論文誌，Vol.44, No.2, pp443-450, 2003