大規模テキスト分類

内山 将夫　　井佐原 均
通信総合研究所

1 はじめに

テキスト分類は日常的な行動である．たとえば，個人としては，メイルやファイルを分類する．また，インターネットにおいても，Yahoo! とか Open Directory Project とかでは，ウェブサイトが分類されている．また，特許や新聞記事なども，カテゴリに分類されている．

そのため，テキスト分類[1] の研究には意義があり，多くの研究が行われている．そして，各種の手法が比較されている [YL99] が，その比較は1万程度の新聞記事による小規模な実験[2] のため，そのときの比較結果がもっと大規模な場合にも，そのまま適用できるのかは不明である．そのため，テキスト分類を大規模に適用したいときには，改めて大規模に実験をした方が良い．

以上の考察に基づき，本稿では，1万記事程度の小規模な場合と100万記事程度の大規模な場合とについて，テキスト分類の実験を行なう．比較した手法は，k-Nearest Neighbor (kNN), 最大エントロピー法 (ME), Naive Bayes (NB) である[3]．

実験の結果として，大規模な場合には，kNN と ME とは，統計的には，kNN が有意に精度が高かったが，その絶対的な差は1ポイント未満であり，実質的に問題となるような精度差ではなかった．そのため，各手法の特徴に応じて使い分けるのが良い．なお，NB は統計的にも実質的にも他の手法より劣っていた．

kNN と ME とを比べたとき，kNN の特徴は，図1に示すように，カテゴリ付与の対象に似た記事が得られることである．これは，インタラクティブにカテゴリを付与するときには，非常に有効である．なぜなら，カテゴリ付与の最終的な確認をユーザがする場合には，カテゴリ付与の根拠を提示する必要があるからである．この根拠として，kNN は類似記事を与えることができるが，ME もしくはその他の学習結果を抽象化して利用する機械学習法では，カテゴリに付与されたスコアしか根拠として与えることができない．両者を比べたとき，根拠としては，ユーザにとっては，類似記事の方が有効であると言える．そのため，インタラクティブにテキスト分類するときには，kNN が適している．

ME の kNN に比べて有利な点は，分類速度が速い点である．kNN の分類に掛る時間は訓練記事数にほぼ比例する．一方，ME の分類速度は，訓練記事数には無関係である．実際，我々の実装においては，大規模な場合について，1記事あたり，kNN では分類に3秒程度，ME では 0.005 秒程度であった．そのため，全自動で分類をする場合には，ME の方が適している．

2 実験データ

大規模な実験には，読売新聞記事データにおいて，日本語記事の 1996 ～ 2000 年の 913118 件を訓練，2001 年 1 ～ 6 月の 169645 件をパラメタ調整，2001 年 7 ～ 12 月の 181863 件をテストに用いた．小規模な場合には，[YL99] と同程度にするために，大規模な場合から無作為抽出した，7796 件を訓練，3017 件をパラメタ調整，3003 件をテストに用いた．

これらの記事には，一記事あたり最大で3つの分類カテゴリが付いている．その種類は，大規模な訓練データでは 75 であり，小規模な場合では 69 である．各記事は，茶筌[4] により形態素解析し，そこから主に (平仮名を除く) 名詞を抽出し，それを機械学習に使う素性とした．その異なり数は，大規模な場合には 285332 であり，小規模な場合には，50043 である．

3 評価方法

評価としては，評価対象の手法が使われる状況に則したものを用いる必要がある．ここでは，テキストに対して n 個のカテゴリを与えるという作業を考え，そ

[1] 本稿では，テキスト分類とは，規定のカテゴリにテキストを割当てることを言う．

[2] [CS02] では，80 万記事程度について比較実験をしている．これは本稿のものとオーダーは同じである．しかし，彼らの実験では，kNN や SVM 等，テキスト分類において良いと認められている方法は実験されていないので，それらの精度を確認することはできない．なお，今後は，1 万記事程度の実験は少なくなり，100 万記事程度の実験が多くなるのではないかと考えられる．

[3] その他に比較したかった方法としては，Support Vector Machine (SVM) があるが，SVM は，大規模な場合には，本稿に間にあうよう (3 週間程度) には学習が終了しそうになかったので，比較からは除いた．なお，従来の結果では，SVM と kNN とは同等な精度である [YL99].

[4] http://chasen.aist-nara.ac.jp/

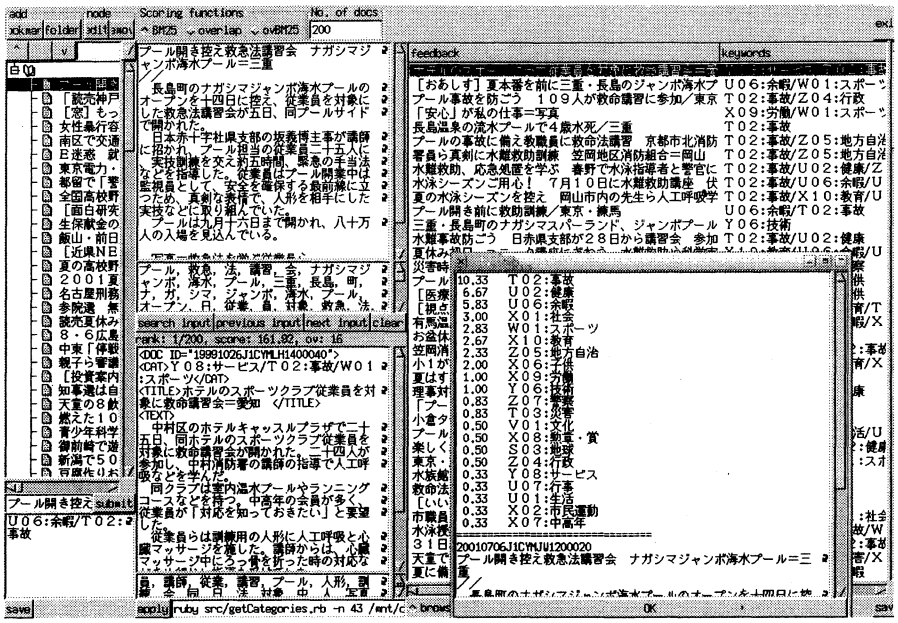

図 1: 中央上窓に質問記事 (分類対象記事) を入力すると右窓に類似記事のタイトルとカテゴリが表示され，選択されたタイトルの本文が中央下窓に表示される．独立表示される窓には kNN によりソートされたカテゴリがある．左端窓は質問記事のリストであり，その下は現在の質問記事の正解カテゴリである．

の作業を補助することを考える．そのためには，各手法がカテゴリをソートしたとき，その上位 n 個に入っている正解の数 c が重要である．ここで，n は各記事について，実際に割当てられているカテゴリの数とすると，$100 * c/n$ は R 精度と呼ばれるものであり，割当てられたカテゴリ数に対する正解のパーセンテージである．R 精度が高いときには，カテゴリ付与をほぼ自動化できると考えられるので，R 精度は，カテゴリ付与の精度の尺度として適当であると考える．この R 精度を各記事毎に求めて，それを平均したものも R 精度と呼ぶ．本稿で精度について述べるときには，この平均された R 精度のことである．

4 比較手法

kNN,ME,NB の順に比較手法の概要を述べる．

まず，kNN では，分類対象の記事 Q に類似した記事 D を k 個集めてきて，そこでのカテゴリの分布に基づきカテゴリにスコアを付ける．ここで，Q との類似度が $r(1 \leq r \leq k)$ 番目に高い記事におけるカテゴリの集合を $C(r)$ とすると，r 番目の記事に付与された各カテゴリ c のスコアを $s(c|r) = 1/|C(r)|$ if $c \in C(r)$ else 0 とし，k 位までのスコアを $S(c) = \sum_{1 \leq r \leq k} s(c|r)$ とする．このスコア $S(c)$ の降順にカテゴリをソートし，その上位 n 個を分類対象のテキストのカテゴリとする．ここで k はパラメタ調整用のデータにおいて最大精度であった k を選ぶが，その値は大規模データでは $k = 43$，小規模データでは $k = 19$ であった．また，k と R 精度との関係はテストデータについて図 2 のようである．図より，大規模データ (Large) については，k を大きくしていっても R 精度はほぼ一定だが，小規模データ (Small) については，R 精度の減衰が大きいことが分かる．これは，データが大きくなれば，類似事例が増えることを端的に表現している．

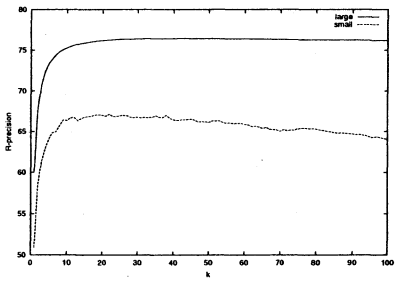

図 2: k と R 精度

Q と D との類似度は BM25[RW00] を変形した以下の式を用いる．変形した理由は，BM25 では Q における形態素の頻度も式に組み込んでいるが，それを組込まない以下の式の方が精度が良かったからである．

$$\sum_{T \in Q} \frac{2tf}{tf + \frac{dl}{avdl}} \log \frac{N - n + 0.5}{n + 0.5}$$

ただし，T は Q に含まれる形態素，tf は D に含まれる T の数，N は検索対象の全記事数，n は T を含む記事の数，dl は D の長さ (延べ形態素数)，$avdl$ は記事の平均長である．

kNN において，類似度の設定，および，選択された k 記事を利用したカテゴリの順位付けについては，様々なバリエーションがある．上述の設定は，類似度は，情報検索において実績のある BM25 に基づき，カテゴリの順位付けのスコアは，パラメタ調整用のデータにおいて，成績の良いものとしたものである．これらは，たとえば，[YL99] のものとは異なる．しかし，これらの詳細な比較は本稿の主眼ではないので，利用した実装 ruby-ir [5] に便利なものとした．

次に，ME では，[NLM99] と同様に Gaussian Prior を用いて各カテゴリの確率を推定した．ME の実装には maxent を用いた．Gaussian Prior における分散の値は，パラメタ調整用のデータにおけるカテゴリの確率分布のエントロピーが極小となる値を用いた．ここでは，ME の詳細は省略するが，特別な手当としては，訓練において，複数カテゴリ m 個が 1 記事に割当てられているときには，各カテゴリが $1/m$ の頻度で，その記事に割当てられているものとして，カテゴリの頻度を計数した．このことは NB の場合も同様である．

各記事の素性としては，形態素を利用したのだが，そのときに，頻度情報を取り入れないと精度が落ちるので，頻度を取り入れる必要がある．そのために，1 回以上出現した形態素，2 回以上出現した形態素，3 回以上出現した形態素について，「形態素表記：1」，「形態素表記：2」，「形態素表記：3」を素性として利用した．そのため，1 記事中に複数回出現している形態素については，複数個の素性が利用されていることになる．たとえば，「音楽」が 2 回出現しているとすると，素性としては，「音楽：1」と「音楽：2」とが利用される．そして，それぞれの素性について，その素性が出現する場合には素性値として 1，しない場合には 0 を与えて確率を推定した．

各カテゴリのソートにおいては，分類対象とする記事中にある素性から各カテゴリの確率を求め，その降順にカテゴリをソートし，その上位 n 位を利用した．これは NB についても同様である．

最後に，NB は，[ELM03] で比較されているものから，最高精度であった多項分布モデルを実装した．

[5] 後述の maxent とともに http://www.crl.go.jp/jt/a132/members/mutiyama/software.html にある．

5 実験結果

R 精度を計算する場合に，今回の記事データの場合には，付与されるカテゴリの数は $n = 1, 2, 3$ のいずれかであるので，そのそれぞれ，および，全体を平均した場合について，R 精度を計算することにする．なお，前掲の図 2 は全体の R 精度についての図である．

小規模な場合の実験結果を表 1，大規模な場合を表 2 に示す．これらの表において「>」は，Welch 検定による片側検定により，有意水準 1%で有意差があることを示し，「~」は有意差がないことを示す．

表 1: 小規模な場合の R 精度

	ME		kNN		NB	数
全体	67.4	~	67.1	>	62.8	3003
n=1	79.8	~	80.5	>	75.2	937
n=2	65.2	~	65.0	>	61.4	1130
n=3	57.7	~	56.1	>	52.1	936

表 2: 大規模な場合の R 精度

	kNN		ME		NB	数
全体	76.44	>	75.96	>	65.78	181863
n=1	86.23	>	84.53	>	74.93	56416
n=2	75.93	~	75.87	>	65.33	70474
n=3	67.05	~	67.27	>	56.98	54973

これらの表より，NB は，その他の手法よりも統計的に有意に劣っているだけでなく，実質的な精度差も，全体について，小規模な場合で約 4 ポイント，大規模な場合で約 10 ポイントであるので，NB は実質的にも劣っている．一方，kNN と ME とは，大規模な場合の全体と $n = 1$ において，統計的に有意差があるが，実質的な差は，全体では，0.5 ポイント程度なので，1 節で述べたように，用途に応じて，これらの手法を使い分けるのが良いと言える．

また，1 節では，「小規模実験の結果が大規模実験に適用できるかは不明である」とも述べたが，そのことについては，今回の実験から言えば，大規模実験をすることにより，より分かることが増える，と言いかえることができる．たとえば，小規模な場合においては，kNN と ME とで有意差がないが，大規模な場合には，有意差がある．そのため，実質的な差がそれほどないとしても，統計的には，kNN の方が優れている．これ

は，大規模な実験をすることにより分かったことである．一般的にいって，大規模な実験をすると，有意差がでやすいので，システムの優劣をはっきりさせたいときには便利である．

なお，素性を拡張した場合として，これまでの実験では，タイトル中の形態素も本文中の形態素も区別せずにいたが，タイトル中の形態素には,「t_」を語頭に付けることにより，全然別のものとして扱い，更に，これまでの実験では除去されていた平仮名の形態素も素性に追加したときの精度を大規模実験の場合について，kNN と ME とについて示すと表 3 のようになる[6]．表 2 と表 3 とを比べると，全体と $n = 1, 2, 3$ の全ての場合について，kNN と ME と共に，表 3 の精度が高く，かつ，その差は片側 Welch 検定で 1%で統計的に有意である．したがって，タイトルもしくは平仮名のどちらかあるいは双方が有効な素性であることが分かる．

表 3: 大規模な場合について拡張した素性での R 精度

	kNN		ME	数
全体	77.48	>	77.22	181863
n=1	87.07	>	85.34	56416
n=2	77.21	∼	77.41	70474
n=3	67.98	<	68.65	54973

このように素性を拡張する動機となったのは，無作為抽出したパラメタ調整用データについて，図 1 のようにインタラクティブなカテゴリ付与をした結果として，これらの情報が重要そうだと考えたからである．たとえば，選挙の記事などでは，タイトルには「候補者に聞く～」のようになっていて選挙のようであっても，その記事内容は，当選後の行政方針であったりして，その結果選挙と分類されなくなっていたりした．また，平仮名を当初除いていた理由は，これまで毎日新聞の情報検索においては平仮名があまり役に立たなかったからであるが，読売新聞においては，平仮名がキーワードとなることがあるようであったので，これを加えた．上記，追加実験では，2 種類の素性を混ぜて追加してしまったので，有効な素性を選択的に調べることはできなかったが，インタラクティブな実験が，素性の吟味に有効であることは分かった．もちろん，素性としてタイトルが重要なことは，ほぼ当然と言えるが，それを当初考慮していない状況から，考えなおして，採用するということはインタラクティブな実験をしたからであろうと思われる．そのため，実際のユーザへのインターフェースとしてだけでなく，高精度なシステムの実現のためにも，特に研究初期の段階では，インタラクティブなテキスト分類は重要であると考える．

その他，インタラクティブな実験をすることにより分かったことをあげると，読売新聞に元から付けられているカテゴリには，付与の際の揺れが多いことである．たとえば，カテゴリ自体にも,「科学」関係と「技術」関係など揺れそうなものがある．また，実際の記事をみても，ほぼ同一内容の記事にも関わらず，異なるカテゴリが付与されているものがある．たとえば，図 1 では,「救急法講習会」についての質問記事 (分類対象記事) においては「余暇／事故」がカテゴリとして与えられているが,「救命講習会」として第 1 位に検索された記事のカテゴリは「サービス／事故／スポーツ」であり，重なっているカテゴリは「事故」のみである．このような揺れは，機械学習の観点からは好ましくなく，この揺れにより，精度に，ある上限があることは確かであろう．

しかし，このカテゴリの違いが実際に揺れであるのか，それとも，正しいカテゴリ付けなのかは，現時点では明かではない．また，このような微妙なカテゴリの付け方が，実応用にとって，どう影響するかも明かではない．このような揺れが存在する場合には，一方では「余暇」から，他方では「スポーツ」から,「救急法講習会」や「救命講習会」を検索できるので，多方面からの検索には，有益である可能性もある．

このように，実際のカテゴリの分布とその分布のもつ (検索などの) 実応用に対しての影響，実際のカテゴリの分布とカテゴリの相関関係から導出できるであろう機械学習手法の精度の上限，あるいは，インタラクティブなテキスト分類の効用など，今後の課題は，分類精度の向上だけに限らず，多い．

[6] kNN については $k = 42$ である．

参考文献

[CS02] Koby Crammer and Yoram Singer. A new family of online algorithms for category ranking. In *SIGIR'2002*, 2002.

[ELM03] Susana Eyheramendy, David D. Lewis, and David Madigan. On the naive Bayes model for text categorization. In *Proc. of AI and Statistics*, 2003.

[NLM99] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *IJCAI Workshop on Machine Learning for Information Filtering*, 1999.

[RW00] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *Proc. of TREC 8*, pages 151–162, 2000.

[YL99] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *SIGIR'99*, 1999.