文書分類のための共クラスタリング

高村大也† 松本裕治†

共クラスタリングを用いて，文書分類の精度を向上させる方法を提案する．文書分類に対する確率モデルによるアプローチでは，同一カテゴリ文書は同一の確率分布から生起したものと仮定されている場合が多い．我々は，そのような仮定が誤りであることを実験的に示し，また共クラスタリングを用いてその問題を軽減する枠組みを提案する．提案手法では，上記仮定が成り立つように訓練文書がクラスタ化され，またデータスパースネス問題を軽減するために，文書を表現している素性も同時にクラスタ化される．この共クラスタリング手法の有効性を示すために実験を行い，精度の向上を確認した．

Co-clustering for Text Categorization

Hiroya Takamura† and Yuji Matsumoto†

We propose a new method to improve the accuracy of Text Categorization using Co-clustering. In a number of previous probabilistic approaches, texts in the same category are implicitly assumed to be generated from an identical distribution. We empirically show that this assumption is not accurate, and propose a new framework based on co-clustering to alleviate this problem. In our method, training texts are clustered so that the assumption is more likely to be true, and at the same time, features are also clustered in order to tackle the data sparseness problem. We conduct some experiments to validate this co-clustering method.

1. はじめに

文書分類は，与えられた文書に対し，その内容に従ってカテゴリを割り当てるタスクである．文書分類に対する多くの確率モデルによるアプローチにおいては，同じカテゴリの文書は同一の分布から生成される（ここでは同一分布の仮定と呼ぶことにする）という仮定を置いている．しかし，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．次章で詳しく論ずるが，この仮定は正しいとはいえず，より正確な確率モデルを求めるための障害になっていると考えられる．また，文書分類における別の問題としては，自然言語処理におけるタスクでよくいわれることであるが，データスパースネス問題がある．これは，素性として単語を用いると素性空間の次元が非常に大きくなるが，各素性の頻度は小さいので信頼性のある統計量を推定することが困難になるという問題である．

これらの問題に対処するため，我々は共クラスタリングに基づいた新しい文書分類の枠組みを提案する．我々の提案手法では，まず文書に対しクラスタリングが行われる．その際，各クラスタの文書は同一の分布から生成したと考えられるようにクラスタが形成される．その後，データスパースネス問題を軽減するために単語のクラスタリングが行われる．つまり，**文書と単語の両方がクラスタリングされる**．その後に一般的な分類器により分類が行われる．

実験により，提案手法が確率的分類器の分類精度を向上させることが示された．

自然言語処理の分野において，クラスタリングは様々な形で応用されてきている[3),8)]．それらの応用の中で，文献 2) では，class-distributional クラスタリングを提案し，彼らの提案手法が Naive Bayes スコアの点で最適なクラスタリングであることを理論的に証明し，また実験的にその有効性を示した．class-distributional クラスタリングにおいては，各単語が与えられたときのカテゴリの生起が確率分布と見なされ，単語はその確率分布に従ってクラスタリングされる．

文献 14) においては，Information Bottleneck 法[15)]が文書分類に適用されている．しかし，文献 2) と 14)

† 奈良先端科学技術大学院大学自然言語処理学講座
Computational Linguistics Laboratories, Nara Institute of Science and Technology

のどちらにおいても，単語のクラスタリングのみが扱われており，単語と文書の共クラスタリングを利用した我々の手法とは異なる．

単語と文書の両方をクラスタリングするという考えそのものは，文献 4)，13) などでも見られるが，それらの研究における目的は文書のクラスタリングであり，文書分類に応用できる枠組みではない．

本稿の構成は以下のとおりである．2 章で簡単な実験を通して，文書分類における同一分布の仮定について考える．3 章で，提案するクラスタリング手法の説明を行う．4 章で分類方法の説明を行う．5 章で実験と結果を示し，それについての考察をする．最後に 6 章で本稿のまとめをする．

2. 人手によるカテゴリと確率的構造

前章で触れたように，人手で定義されたカテゴリは確率的構造に基づいていないと思われる．これを実験的に調査するために，我々は Reuters-21578 データセットを用いて小規模な実験を行った．まず “earn” というカテゴリにラベル付けされたテキストをクラスタリングし，7 個のクラスタが得られた☆．これらのクラスタを，クラスタ A，クラスタ B，···，クラスタ G と呼んでおく．各クラスタのサイズは極端に小さいものではない（100 以上の文書を含む）．ここで，クラスタあるいはカテゴリが与えられたときの単語の生起を確率事象と見なし，“earn” の 1 つのクラスタ A と他のクラスタ間の KL（Kullback-Leibler，カルバック–ライブラー）ダイバージェンス，およびそのクラスタ A と（頻度の大きい方から 10 個の）各カテゴリ間の KL ダイバージェンスを計算した．

もし，同一カテゴリに属する文書における単語生起が（近似的にでも）同一分布の仮定に従えば，“earn” のクラスタ間の KL ダイバージェンスは，“earn” のクラスタ A と他のカテゴリ間の KL ダイバージェンスと比較して，小さくなることが予想される．結果を**図 1** に示す．この図においては，クラスタ A からの KL ダイバージェンスが示されている．点線で表された左から 6 個の長方形は，“earn” の A 以外のクラスタに対応し，中央の長方形は “earn” に対応し，右側の 9 個の長方形は他のカテゴリに対応している．KL ダイバージェンスは，“earn” の 2 つのクラスタに関し，非常に大きく，さきほどの予想とかけ離れている．これだけでは不十分なので，さらにクラスタ A から

☆ クラスタ数は AIC（Akaike Information Criterion，赤池情報量規準）によって決定された．AIC の適用法の詳細は後で説明する．

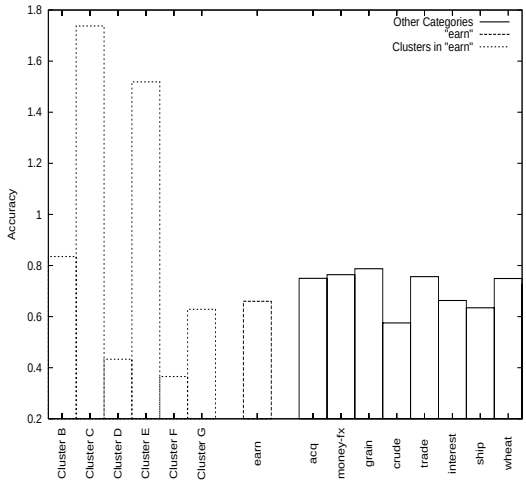

図 1 KL ダイバージェンス
Fig. 1 KL-divergence.

“earn” へのダイバージェンスと，クラスタ A から他のカテゴリへのダイバージェンスを比較してみる．すると，“earn” へのダイバージェンスより小さなダイバージェンスを持つようなカテゴリが 2 つ存在することが分かる．“earn” の中に，“earn” と比較して他のカテゴリの方に近い文書群（クラスタ）が存在することを示している．

頻度の大きい方から 10 個のカテゴリについてもクラスタリングを施し同様な実験を行った結果，他のカテゴリに近いクラスタが存在する現象は，クラスタリング後も複数のクラスタを含む 8 カテゴリ（“earn” も含む）中の 3 カテゴリで見られた．

この実験は小規模であるが，同一分布の仮定がつねに成り立つわけではないことを示唆している．また，同一分布の仮定がより妥当になるように文書をクラスタリングすることによって，より精密な分類モデルが得られる可能性があるといえよう．

3. 文書と単語の共クラスタリング

この章では，同一分布の仮定の非妥当性より生じる問題を解決する枠組みを提案する．

我々のアプローチはボトムアップのクラスタリング手法に基づいている．初期段階では，各クラスタは 1 つの単語もしくは文書からなる．各ステップにおいて，選ばれた 2 つの単語クラスタ，もしくは 2 つの文書クラスタが 1 つのクラスタにマージされる．マージするときの基準として用いるのは，マージによって生じる尤度減少である．この基準は，相互情報量やジェンセン=シャノン・ダイバージェンスと関係がある．以下

に，我々の用いる確率モデルを詳しく説明し，さらにクラスタリングのアルゴリズム，マージの基準における背景，クラスタリングの停止条件について述べる．

3.1 確率モデル

我々が用いる確率モデルは，文献 8) の hard クラスタリングモデルで使用されているモデルである．文献 8) では，2 つの単語の共起が取り扱われているが，我々が欲しいのは単語と文書の共起である．我々のモデルは次のように表される：

$$P(w,d) = P(C_w,C_d)P(w|C_w)P(d|C_d) \quad (1)$$
$$w \in C_w, d \in C_d$$

ここで，w はある単語，d はある文書を表す．また，C_w と C_d はそれぞれ w と d が属するクラスタを表す．

単語と文書の共起データ：

$$S = \{(w_1,d_1),(w_2,d_2),\cdots,(w_m,d_m)\}, \quad (2)$$

が与えられたとき，その対数尤度は，

$$\sum_{(w,d)\in S} \log P(w,d)$$
$$= \sum_{(w,d)\in S} \log P(C_w,C_d)P(w|C_w)P(d|C_d) \quad (3)$$

と計算される．モデル (1) のパラメータは，最大尤度推定に基づいて，

$$P(C_w,C_d) = \frac{N(C_w,C_d)}{|S|}, \quad (4)$$

$$P(w|C_w) = \frac{N(w)}{N(C_w)}, \quad (5)$$

$$P(d|C_d) = \frac{N(d)}{N(C_d)}, \quad (6)$$

と計算される．ここで，$N(x)$ は x の頻度を指す．

3.2 クラスタリング・アルゴリズム

文献 8) で記述されているアルゴリズムでは，2 つの正整数 k と l が与えられ，第 1 次元のマージが k 回行われ，続いて第 2 次元のマージが l 回行われるというプロセスが繰り返される．

我々は，文献 8) のアルゴリズムと異なる，2 つのアルゴリズムを提案する．この 2 つのどちらも，3.1 節で述べた確率モデルをベースにしている．どちらのアルゴリズムにおいても，各ステップにおいて，ある単語クラスタのペアもしくは文書クラスタのペアがマージされる．両アルゴリズムの差異は，そのペアの選び方である．1 つは，文書クラスタリングだけが連続して（停止条件が成り立つまで）先に行われ，その後単語クラスタリングが行われるもので，これを *text-first* クラスタリングと呼ぶことにする．もう 1 つは，各ステップで，すべての単語クラスタのペアと文書クラスタのペアの中から，尤度減少が最小のものを選び，それをマージするものである．これをここでは *greedy* クラスタリングと呼ぶことにする．つまり pseudo コードで表すと，

- Text-first クラスタリング
  - (1) 初期化
  - (2) 停止条件が成立するまで，尤度減少最小の文書クラスタペアのマージを繰り返す．
  - (3) 停止条件が成立するまで，尤度減少最小の単語クラスタペアのマージを繰り返す．
- Greedy クラスタリング
  - (1) 初期化
  - (2) 停止条件が成立するまで，尤度減少最小の文書もしくは単語クラスタペアのマージを繰り返す．

ということになる．マージされるペアの選択に際しては，我々は 2 つの制約を置いた．1 つは，文書のマージは同一カテゴリ内で行われなければならないこと，もう 1 つは単語のマージは同一品詞内で行われなければならないことである．第 1 の制約は，後で述べる我々の分類手法にとって欠くことのできないものである．両制約ともに探索空間を小さくするので，クラスタリングにかかる計算時間を短くする効果がある．

text-first クラスタリングは，class-distribution の情報を使って単語クラスタリングが行われるという利点がある☆．class-distributional クラスタリングは，text-first クラスタリングの特別な場合と見なすこともできる．すなわち，もし文書クラスタリングステップの停止条件を “制約を冒さずにクラスタをマージできない” とすれば，text-first クラスタリングは，class-distributional クラスタリングと同一のものとなる．

3.3 尤度減少最小基準と他の基準との関係

クラスタリングのそもそもの目的が，同一と考えられる確率分布を持つクラスタを形成することであったにもかかわらず，尤度減少最小基準を用いることは不思議に感じられるかもしれない．しかし，尤度減少最小基準は，確率分布間の距離としてある尺度を用いて，その距離が最小となるものを選ぶことと同値であることをここで示す．ΔL は，単語 i と j をマージすることにより起こる対数尤度 (3) の減少を表すものとする．また，$|S|$ で全訓練データ事例数を表すものとする．クラスタどうしは互いに排反であることから，$P(C_{ij},C_d) = P(C_i,C_d) + P(C_j,C_d)$ が成り立つこ

☆ 厳密には，使われる情報は class-distribution そのものとは異なる．しかし，文書クラスタリングがある程度進むと，class-distribution に近い情報が利用できる．

とを利用すると，

$$\begin{aligned}\frac{\Delta L}{|S|} &= \sum_{C_d} -P(C_{ij},C_d)\log\frac{P(C_{ij},C_d)}{P(C_{ij})P(C_d)} \\ &\quad + \sum_{C_d} P(C_i,C_d)\log\frac{P(C_i,C_d)}{P(C_i)P(C_d)} \\ &\quad + \sum_{C_d} P(C_j,C_d)\log\frac{P(C_j,C_d)}{P(C_j)P(C_d)} \\ &= \sum_{C_d} P(C_i,C_d)\{\log\frac{P(C_i,C_d)}{P(C_i)P(C_d)} \\ &\quad - \log\frac{P(C_{ij},C_d)}{P(C_{ij})P(C_d)}\} \\ &\quad + \sum_{C_d} P(C_j,C_d)\{\log\frac{P(C_j,C_d)}{P(C_j)P(C_d)} \\ &\quad - \log\frac{P(C_{ij},C_d)}{P(C_{ij})P(C_d)}\} \\ &= P(C_i)\sum_{C_d} P(C_d|C_i)\log\frac{P(C_d|C_i)}{P(C_d|C_{ij})} \\ &\quad + P(C_j)\sum_{C_d} P(C_d|C_j)\log\frac{P(C_d|C_j)}{P(C_d|C_{ij})} \\ &= P(C_i)D_{KL}(P(\cdot|C_i)||P(\cdot|C_{ij})) \\ &\quad + P(C_j)D_{KL}(P(\cdot|C_j)||P(\cdot|C_{ij})) \quad (7)\end{aligned}$$

と変形できる．ここで $D_{KL}(p||q)$ は確率分布 p と q の KL ダイバージェンスを表す．さて，式 (7) は，文献 2) で，平均化 KL ダイバージェンス（KL divergence to the mean）として使われているものである．つまり，尤度減少最小基準を用いることは，平均化 KL ダイバージェンスの意味で最も近いクラスタをマージしていくことと同値であることが分かる．逆にいえば，文献 2) で用いられているクラスタリング手法は，尤度減少最小という点で妥当性があるといえる．

また，文献 8) などでは，相互情報量との関係が指摘されている．

3.4 AIC を用いた停止条件

クラスタリングの停止条件としては，AIC（Akaike Information Criterion，赤池情報量規準）[1] を用いた．類似したアルゴリズムを提案している文献 8) では，MDL（Minimum Description Length，最小記述長）原理[11] を用いているが，我々は MDL は用いなかった．理由は，MDL は予備実験において小さすぎるクラスタ数を予測する傾向があったからである．文書クラスタリングにおいては，MDL はカテゴリ数よりも小さなクラスタ数を予測したが，我々は異なるカテゴリに属するクラスタどうしはマージできないという仮定を置いているため，そのようなクラスタ数は我々の手法に合わない．つまり，そのような状況においては 1 つのカテゴリが 1 つのクラスタを成すことになり，文書クラスタリングを設定した影響がなくなってしまう．

AIC は以下のように応用される．マージによって引き起こされるパラメータ数の減少は：

$$\Delta N_p = \begin{cases} |C(\mathbf{D})| - 1, & \text{(word-merge)} \\ |C(\mathbf{W})| - 1, & \text{(text-merge)} \end{cases} \quad (8)$$

と表される．ここで，

$|C(\mathbf{D})|$ = Number of word-clusters,
$|C(\mathbf{W})|$ = Number of text-clusters.

である．AIC に従うと，停止条件は，

$$-\Delta L + \Delta N_p > 0. \quad (9)$$

となる．ここで ΔL は，マージにより起こる対数尤度 (3) の減少を表す．

text-first クラスタリングにおいては，AIC が適用される点が 2 つあることに注意されたい．1 つは，文書クラスタリングステップの終了点であり，もう 1 つは単語クラスタリングステップの終了点である．

4. 分　　類

確率モデルに基づいた分類器に関しては，提案するクラスタリング手法と組み合わせて，精度向上が期待されるが，確率モデルに基づいていない分類器に関しては予測し難い．よってここでは，その両方の振舞いを観察したいので，2 種類の分類器を用いて実験を行うことにする．それらの分類器に関して簡単に説明を行う．確率モデルに基づいたものとしては，代表的な分類器である NB（Naive Bayes，ナイーブ・ベイズ）分類器[10] を，確率モデルに基づいていないものとしては，SVM（Support Vector Machines，サポート・ベクター・マシン）[16] を用いた．

NB 分類器に関しては，多項分布モデル[9] を採用したが，文書の長さによる影響は無視した．

SVM は，Structural Risk Minimization [16] という考え方を背景に持つ二値分類器である．SVM は高い汎化性能を持ち，文書分類において非常に良い結果を残している[7]．SVM を多値分類に応用するため，我々は one-versus-rest 方式を用いた．この方式においては，各カテゴリに関し，そのカテゴリか否かを表す二値分類器が構成される．分類段階では，与えられた事例について，各二値分類器がそのカテゴリを支持する度合い（SVM では分離平面からの距離）が最も大きいカテゴリに属するものと判定される．

我々の手法においては，文書クラスタリングが前もってされている．よって，まずテスト文書を分類し，各テスト文書がどのクラスタに属するかを予測する．次に，各テスト文書に対し，それが属すると予測された

表 1 Greedy クラスタリング
Table 1 Greedy clustering.

単語圧縮率 (%)	100.0	90.0	80.0	70.0	60.0	50.0	40.0	30.0	20.0
文書圧縮率 (%)	100.0	94.8	94.3	93.8	93.1	91.8	43.1	29.9	17.0
精度	0.717	0.718	0.719	0.722	0.731	0.739	0.807	0.834	0.836

10.0	9.7 (AIC)	9.0	8.0	7.0	6.0	5.0	4.0	3.0	2.0
7.1	6.8	6.2	5.5	4.9	4.1	3.5	2.8	2.3	1.8
0.848	0.848	0.847	0.848	0.849	0.846	0.846	0.843	0.837	0.841

クラスタのカテゴリを付与する（提案手法では，各クラスタ内の文書はすべて同一のカテゴリタグを持つ）．ただし，SVM に関しては，二値分類器の特性を考慮して，学習段階で同一カテゴリで他クラスタに属する文書を取り除いた．

5. 実　　験

5.1 実験設定

実験に用いたデータセットは，Reuters-21578☆である．データセットから訓練文書とテスト文書を抽出する 1 つの方法である ModApte-split[6] を実行した後，さらに本文が無意味と思われるような文書を削除した．最終的に，8,815 訓練文書と 3,023 テスト文書が残り，カテゴリの種類数は 116 となった．素性として使用した単語は，名詞，動詞，固有名詞，形容詞，および副詞のうち，全訓練データ中で 5 回以上出現したものである．ステミングは，TreeTagger[12] を用いて行った．

複数のラベルを持つ訓練文書については，そのような文書を付与されているラベル数だけ複製し，各複製文書に 1 つのラベルを付けてクラスタリングを行った．クラスタリング後，複製した文書を 1 つに戻し，この文書は複製文書が属していたクラスタすべてに属するものとして扱った．分類のためのクラスタリングとして性能が認められている class-distributional クラスタリングとの比較を行った．

SVM による分類に際しては，TinySVM☆☆を使用した．また，SVM の学習において用いたカーネル関数は，線形カーネルである．

評価は，両分類器とも，分類精度を用いた．ここでいう分類精度とは，全テスト文書中の正しく分類できた文書の割合である．また，複数のラベルが付いているテスト事例に関しては，予測されたカテゴリが答えのラベルに含まれている場合を正解とした．事例が複数のラベルを持つ場合の評価値としては，一般的にはカテゴリごとの F 値や break-even ポイントなどの方がよく使用されるが，NB 分類器は二値分類に向かないので，NB 分類器と SVM の分類性能が直接比較できるようにここでは上記分類精度を用いた．詳しく述べると，NB 分類器はその分類性能は認められているにもかかわらず，素性間の独立性という実際のデータでは一般に成立しない仮定を置いているので，算出される確率値自体は信頼できない[5]．つまり，事例について各カテゴリの生成確率を一度に求めるという，NB 分類器の一般的な適用方法を用いた場合，算出される確率値とある閾値の大小に従い各カテゴリについて正例であるか負例であるかを判断することは適切でない．また，訓練事例があるカテゴリに属するか否かに従い二値分類の学習を行うことも技術的には可能であるが，NB 分類器に関しては非常に不利である（正例と負例の数が大きく偏ることが原因と思われる）．文献 7) においても，正例数が小さいカテゴリに関して NB 分類器の break-even ポイントが低下する傾向が現れている．

5.2 実験結果

クラスタリングを行わない状態での NB 分類器および SVM での分類精度は，それぞれ 0.863，0.890 であった．

text-first クラスタリングにおいて，AIC による停止条件に従い，文書クラスタ数として 141 が選ばれた．これはもとのカテゴリ数である 116 と比較してわずかに大きい値である．複数のクラスタを含むカテゴリは，“earn（7 クラスタ）”，“acq（6 クラスタ）”，“others（8 クラスタ）”，“crude（3 クラスタ）”，“money-fx（3 クラスタ）”，“grain（2 クラスタ）”，“interest（2 クラスタ）”，“trade（2 クラスタ）” の 8 カテゴリである．

以下に分類実験の結果を示す．**表 1** は，NB 分類器と，text-first クラスタリングあるいは class-distributional クラスタリングを組み合わせたモデルの分類結果である．結果は様々な圧縮率について算出した．ただし，圧縮率は単語に関する値であり，text-first クラスタリングに関しては，文書クラスタリングステップ終了後の単語クラスタリングステップについての分類結果

☆ http://www.daviddlewis.com/resources/より入手可能.
☆☆ http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/より入手可能.

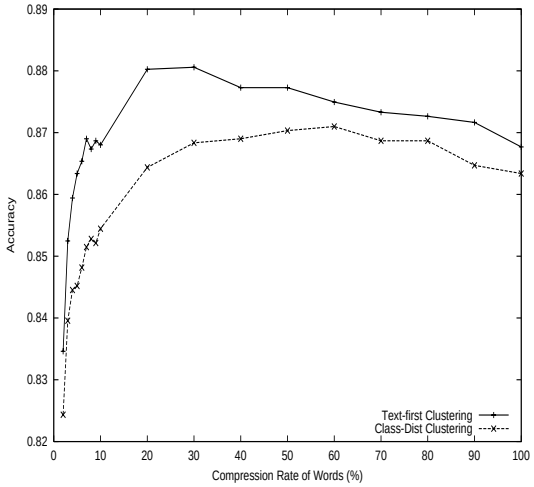

図 2 NB 分類器による分類精度
Fig. 2 Categorization accuracy with NB classifier.

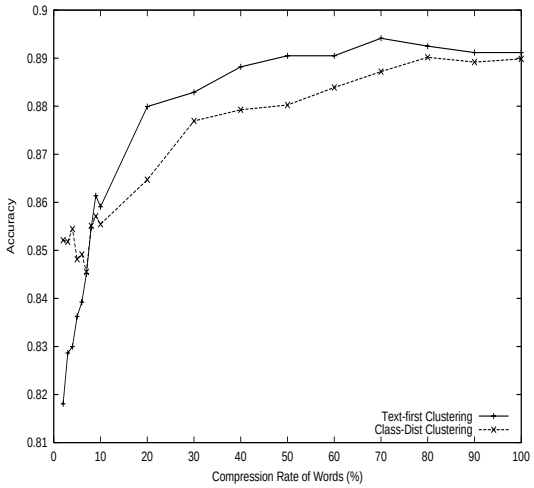

図 3 SVM による分類精度
Fig. 3 Categorization accuracy with Support Vector Machine.

が載せられている．図中，“Text-first Clustering” は提案手法に対応し，“Class-Dist Clustering” は class-distributional クラスタリングに対応する（このような記法は他の図や表でも用いる）．

図 3 は，SVM と，text-first クラスタリングあるいは class-distributional クラスタリングを組み合わせたモデルの分類結果である．

表 2 には，NB 分類器に関して，AIC によって予測された圧縮率と対応する分類精度，および実際の最適圧縮率と対応する分類精度を示した．

表 1 には，greedy クラスタリングと NB 分類器の組合せによる分類の結果を示す．greedy クラスタリングの場合，単語の圧縮率と文書の圧縮率の両方の値が必要なため，**図 2** には含めなかった．また，表 1 には，AIC により予測された最適圧縮率（単語圧縮率は 9.7%，文書圧縮率は 6.8%）と対応する分類精度も記入されている．

表 2 予測された最適圧縮率
Table 2 Prediction of the best compression rate.

クラスタリング手法	圧縮率 (%)	精度
Class-Dist (AIC)	13.4	0.859
(Actual)	60	0.871
Text-first (AIC)	16.6	0.880
(Actual)	30	0.881

5.3 考　　察

まず，図 2 を見てみる．単語圧縮率 100%での値を比較すると，text-first クラスタリングの方が精度が良い．これは，NB 分類器に対し文書クラスタリングの効果が現れていること意味する．つまり，文書カテゴリを細分化したことによる確率モデルの是正の効果である．また，単語圧縮率を下げていくことによって精度が向上している．これは，単語クラスタリングの効果，つまりデータスパースネス問題の軽減による効果を示す．さらに，単語圧縮率を下げていくことによって，両クラスタリング手法の精度の差は大きくなっていくのが分かる．これは，当初の狙いであった共クラスタリングによる効果が現れていることを意味する．

また，図 3 を見ると，様々な単語圧縮率において，text-first クラスタリングが class-distributional クラスタリングを精度において上回っていることが分かる．単語圧縮率 70%付近では text-first クラスタリングは単語圧縮率 100%の精度をわずかに上回っているが，特に小さな単語圧縮率においては精度低下が見られる．class-distributional クラスタリングに関しては，精度向上はまったく見られない．これは，SVM に対しては単語のクラスタリングが効果的とはいえないことを示している．

表 2 に示した AIC によって予測された圧縮率は，実際の最適圧縮率にそれほど近くないものの，text-first クラスタリングに関しては，精度の違いはそれほど大きくない．また，表 2 において AIC によって予測された圧縮率における精度に関し，有意水準 1%で符号検定を行った結果，精度の差は有意であるという結果が得られた．

表 1 に示される greedy クラスタリングは，クラスタリングに時間がかかる割に良い精度が得られていない．これは，クラスタリングの初期の時点において，

class-distribution と無関係に単語のクラスタリングが進んでしまうことが原因ではないかと思われる．

6. おわりに

本稿では，共クラスタリングを利用して文書分類の精度を向上させる方法を提案した．提案手法では，訓練文書と素性の両方がクラスタリングされる．

多くの既存の確率的アプローチが，各カテゴリが 1 つの確率分布を持つという仮定の上に成り立っているが，その仮定が間違っていることを実験的に示した．提案手法は，この仮定の誤りを回避し，同時にデータスパースネス問題も軽減する．

実験により，提案手法は NB 分類器の分類精度を向上させることが示された．また，SVM に対しては単語クラスタリングは精度の悪化を招くが，文書クラスタリングは分類精度を向上させることが示された．

今後の発展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，他のデータセットでの実験も必要と思われる．

また，停止条件として AIC を用いたが，これが妥当であるかは詳しく調査していない．特に，text-first クラスタリングの文書クラスタリングステップの停止条件としては，AIC が適当であるか否かは未知である．この部分にさらなる調査が必要である．また別の基準も試す必要がある．

クラスタリングとしては，ボトムアップ手法を用いたが，トップダウンのクラスタリングも試してみる必要があろう．データによっては，計算時間の大幅な節約につながると思われる．

参考文献

1) Akaike, H.: A New Look at the Statistical Model Identification, *IEEE Trans. Automatic Control*, Vol.19, No.6, pp.716–723 (1974).
2) Baker, D. and McCallum, A.: Distributional Clustering of Words for Text Classification, *Proc. 21st ACM International Conference on Research and Development in Information Retrieval* (*SIGIR98*), pp.96–103 (1998).
3) Brown, P., Pietra, V., deSouza, P., Lai, J. and Mercer, R.: Class-based N-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467–479 (1992).
4) Dhillon, I.: Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning, Technical Report 2001-05, UT Austin CS Department (2001).
5) Domingos, P. and Pazzani, M.: Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier, *Proc. 13th International Conference on Machine Learning*, pp.105–112 (1996).
6) Dumais, S., Platt, J. Heckerman, D. and Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization, *Proc. 7th International Conference on Information and Knowledge Management* (*ACM-CIKM98*), pp.148–155 (1998).
7) Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proc. European Conference on Machine Learning*, pp.137–142 (1998).
8) Li, H. and Abe, N.: Word Clustering and Disambiguation Based on Co-occurrence Data, *Proc. 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics* (*COLING-ACL '98*), pp.749–755 (1998).
9) McCallum, A. and Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification, *Proc. AAAI-98 Workshop on Learning for Text Categorization*, pp.41–48 (1998).
10) Mitchell, T.: *Machine Learning*, McGraw Hill (1997).
11) Rissanen, J.: Stochastic Complexity, *Journal of Royal Statistical Society, Series B*, Vol.49, No.3, pp.223–239 (1987).
12) Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proc. International Conference on New Methods in Language Processing*, pp.44–49 (1994).
13) Slonim, N. and Tishby, N.: Document Clustering using Word Clusters via the Information Bottleneck Method, *Proc. 23th ACM International Conference Research and Development in Information Retrieval* (*SIGIR2000*), pp.208–215 (2000).
14) Slonim, N. and Tishby, N.: The Power of Word Clusters for Text Classification, *23rd European Colloquium on Information Retrieval Research* (2001).
15) Tishby, N., Pereira, F. and Bialek, W.: The Information Bottleneck Method, *Proc. 37th Annual Allerton Conference on Communication, Control and Computing*, pp.368–377 (1999).
16) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag Berlin (1995).

(平成 14 年 5 月 1 日受付)
(平成 14 年 12 月 3 日採録)

高村　大也（学生会員）

1974 年生．1997 年東京大学工学部計数工学科卒業．2000 年同大学大学院工学系研究科計数工学専攻修了（1999 年はオーストリアウィーン工科大学にて研究）．2000 年奈良先端科学技術大学院大学博士後期課程入学，現在に至る．自然言語処理，特に学習理論等の応用に興味を持つ．

松本　裕治（正会員）

1955 年生．1977 年京都大学工学部情報工学科卒業．1979 年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984〜85 年英国インペリアルカレッジ客員研究員．1985〜87 年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993 年より奈良先端科学技術大学院大学教授，現在に至る．京都大学工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM 各会員．