文書分類への二次元クラスタリングの適用

高村 大也　　松本 裕治
奈良先端科学技術大学院大学
情報科学研究科自然言語処理学講座
〒 630-0101 奈良県生駒市高山町 8916-5
0743-72-5246,5248
{hiroya-t,matsu}@is.aist-nara.ac.jp

二次元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, 同一カテゴリー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, またその問題を軽減する枠組を提案する. 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ー化され, 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ー化される. この二次元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.

キーワード : 文書分類, 二次元クラスタリング, サポートベクターマシン, ナイーブ・ベイズ分類器

Two-dimensional Clustering for Text Categorization

Hiroya Takamura　　Yuji Matsumoto
Nara Institute of Science and Technology
Graduate School of Information Science
8916-5 Takamaya, Ikoma, Nara 630-0101, JAPAN
+81-743-72-5246, 5248
{hiroya-t,matsu}@is.aist-nara.ac.jp

We propose a new method to improve the accuracy of Text Categorization using Two-dimensional Clustering. In most of the previous probabilistic approaches, texts in the same category are implicitly assumed to be generated from an identical distribution. We show empirically that this assumption is violated and propose a new framework to alleviate this problem. In our method, training texts are clustered so that i.i.d. 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 two-dimensional clustering method.

Keywords : Text Categorization, Two-dimensional Clustering, Support Vector Machine, Naive Bayes Classifier

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ー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, 同じカテゴリーの文書は i.i.d.（独立に同一の分布から生成される）であるという仮定を置いている. しかし, カテゴリー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言えず, 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言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, 我々は二次元クラスタリングに基づいた新しい文書分類の枠組を提案する. 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, 各クラスター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ーが形成される. 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 (Brown, 1992; Li and Abe, 1998). それらの応用の中で, 文献 (Baker et al, 1998) 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ーの生起が確率分布とみなされ, 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 (Slonim and Tishby, 2001) においては, Information Bottleneck 法 (Tishby et al, 1999) が文書分類に適用されている. しかし, 文献 (Baker et al, 1998) と (Slonim and Tishby, 2001) 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, 単語と文書の二次元クラスタリングを利用した我々の手法とは異なる.

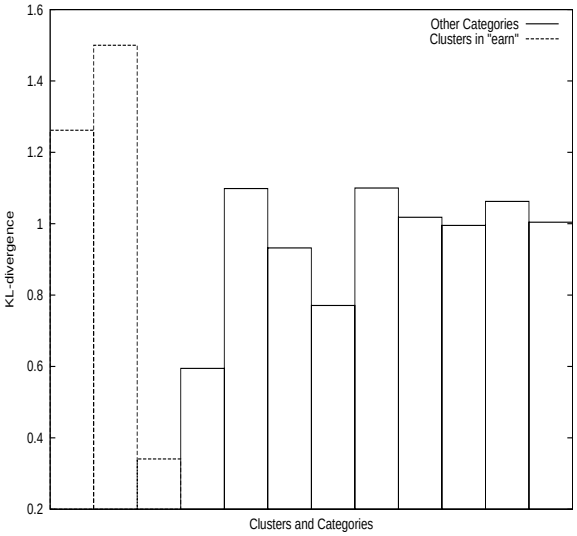

図 1: KL ダイバージェンス

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 (Slonim and Tishby, 2000; Dhillon, 2001) 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そのものであり, 文書分類に応用できる枠組ではない.

本稿の構成は以下の通りである. 第二節で簡単な実験を通して, 文書分類における i.i.d. 仮定について考える. 第三節で, 提案するクラスタリング手法の説明を行う. 第四節で分類方法の説明を行う. 第五節で実験と結果を示し, それについての考察をする. 最後に第六節で本稿のまとめをする.

2 人手によるカテゴリーと確率的構造

前章で触れたように, 人手で定義されたカテゴリーは確率的構造に基づいていないと思われる. これを実験的に示すために, 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” というカテゴリーにラベル付けされたテキストをクラスタリングし, 四つのクラスターが得られた[1]. 各

[1] クラスター数は AIC（Akaike Information Criterion, 赤池情報量規準）によって決定された. AIC の

クラスターのサイズは十分に大きいもの（約数百以上の文書を含む大きさ）である．ここで, クラスターあるいはカテゴリーが与えられたときの単語の生起を確率事象とみなし, “earn” の一つのクラスターと他のクラスター間の KL（Kullback-Leibler, カルバック-ライブラー）ダイバージェンス, 及びその一つのクラスターと他の（頻度の大きい方から 10 個の）カテゴリー間の KL ダイバージェンスを計算した．

もし, 同一カテゴリーに属する文書における単語生起が（近似的にでも）i.i.d. ならば, “earn” のクラスター間の KL ダイバージェンスは, “earn” のクラスターと他のカテゴリー間の 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．結果は図 1 に示す．この図 1 においては, 任意に選ばれたクラスターからの 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三つの長方形は, “earn” の（選ばれていない）他のクラスターに対応し, それ以外の長方形は他のカテゴリーに対応している．KL ダイバージェンスは, “earn” の二つのクラスターに関し, 非常に大きく, さきほどの予想とかけ離れている．

この実験は小規模であるが, i.i.d. の仮定が常に成り立つわけではないことを示唆している．また, i.i.d. 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 文書と単語の二次元クラスタリング

この節では, i.i.d. の仮定の非妥当性より生じる問題を解決する枠組を提案する．

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, 各クラスターは一つの単語もしくは文書から成る．各ステップにおいて, 選ばれた二つの単語クラスター, もしくは二つの文書クラスターが一つのクラスターにマージされる．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 (Li and Abe, 1998) 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 (Li and Abe, 1998) では, 二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:

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

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 が属するクラスターを表す．

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 (Li and Abe, 1998) で記述されているアルゴリズムでは, 二つの正整数 k と l が与えられ, 第一次元のマージが k 回行われ, 続いて第二次元のマージが 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 (Li and Abe, 1998) のアルゴリズムと異なる, 二つのアルゴリズムを提案

適用法の詳細は後で説明する．

する．この二つのどちらも, 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, ある単語クラスターのペアもしくは文書クラスター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．一つは, 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．もう一つは, 各ステップで, 全ての単語クラスターのペアと文書クラスターのペアの中から, 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ー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ー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ー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, 我々は二つの制約を置いた．一つは, 文書のマージは同一カテゴリー内で行われなければならないこと, もう一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．第一の制約は, 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．両制約共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[2]．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みなすこともできる．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ーをマージできない” 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．

[2]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．

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ーを形成することであったにも関わらず, 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．クラスター同士は互いに排反であることから, $P(C_{ij}, C_d) = P(C_i, C_d) + P(C_j, C_d)$ が成り立つことを利用すると,

$$
\begin{aligned}
& \tfrac{1}{|S|}\Delta L \\
= \; & \textstyle\sum_{C_d} -P(C_{ij}, C_d) \log \frac{P(C_{ij}, C_d)}{P(C_{ij})P(C_d)} \\
& + \textstyle\sum_{C_d} P(C_i, C_d) \log \frac{P(C_i, C_d)}{P(C_i)P(C_d)} \\
& + \textstyle\sum_{C_d} P(C_j, C_d) \log \frac{P(C_j, C_d)}{P(C_j)P(C_d)} \\
= \; & \textstyle\sum_{C_d} P(C_i, C_d)\{\log \frac{P(C_i, C_d)}{P(C_i)P(C_d)} \\
& \qquad - \log \frac{P(C_{ij}, C_d)}{P(C_{ij})P(C_d)}\} \\
& + \textstyle\sum_{C_d} P(C_j, C_d)\{\log \frac{P(C_j, C_d)}{P(C_j)P(C_d)} \\
& \qquad - \log \frac{P(C_{ij}, C_d)}{P(C_{ij})P(C_d)}\} \\
= \; & P(C_i) \textstyle\sum_{C_d} P(C_d|C_i) \log \frac{P(C_d|C_i)}{P(C_d|C_{ij})} \\
& + P(C_j) \textstyle\sum_{C_d} P(C_d|C_j) \log \frac{P(C_d|C_j)}{P(C_d|C_{ij})} \\
= \; & P(C_i) D_{KL}(P(\cdot|C_i)||P(\cdot|C_{ij})) \\
& + P(C_j) D_{KL}(P(\cdot|C_j)||P(\cdot|C_{ij})) \qquad (7)
\end{aligned}
$$

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 (Baker et al, 1998) 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ーをマージしていくことと同値であ

ることがわかる. 逆にいえば, 文献 (Baker et al, 1998) 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 (Li and Abe, 1998) 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) (Akaike, 1974) を用いた. 類似したアルゴリズムを提案している文献 (Li and Abe, 1998) では, MDL (Minimum Description Length, 最小記述長) 原理 (Rissanen, 1987) を用いているが, 我々は MDL は用いなかった. 理由は, 予備実験において小さ過ぎるクラスター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, カテゴリー数よりも小さなクラスター数を予測したが, 我々は異なるカテゴリーに属するクラスター同士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ー数は我々の手法に合わない).

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,

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

である. AIC に従うと, 停止条件は,

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

となる.

text-first クラスタリングにおいては, AIC が適用される点が二つあることに注意されたい. 一つは, 文書クラスタリングステップの終了点であり, もう一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, 二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 (Mitchell, 1997) 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) (Vapnik, 1995) を用いた.

NB 分類器に関しては, 多項分布モデル (McCallum and Nigam, 1998) 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(Vapnik, 1995) 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 (Joachims, 1998).

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ー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クラスターのカテゴリーを付与する (提案手法では, 各クラスター内の文書は全て同一のカテゴリータグを持つ). 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ーで他クラスター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[3]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一つの方法である ModApte-split 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, 8815 訓練文書と 3023 テスト文書が残り, カテゴリーの種類数は 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, 及び副詞のうち, 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 (Schmid, 1994) を用いて行った.

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.

[3]http://www.research.att.com/ lewis/ において入手可能.

SVMによる分類に際しては, TinySVM[4]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, 両分類器共, 分類精度を用いた.

5.2 実験結果

クラスタリングを行わない状態でのNB分類器及び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, 文書クラスター数として141が選ばれた. これはもとのカテゴリー数である116と比較して僅かに大きい値である. 複数のクラスターを含むカテゴリーは, “earn（7クラスター）”, “acq（6クラスター）”, “others（8クラスター）”’, “crude（3クラスター）”, “money-fx（3クラスター）”, “grain（2クラスター）”, “interest（2クラスター）”, “trade（2クラスター）”の8カテゴリーである.

以下に分類実験の結果を示す. 図2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1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, 及び実際の最適圧縮率と対応する分類精度を示した.

表2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2には, AICにより予測された最適圧縮率と対応する分類精度も記入されている.

[4]http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/ より入手可能.

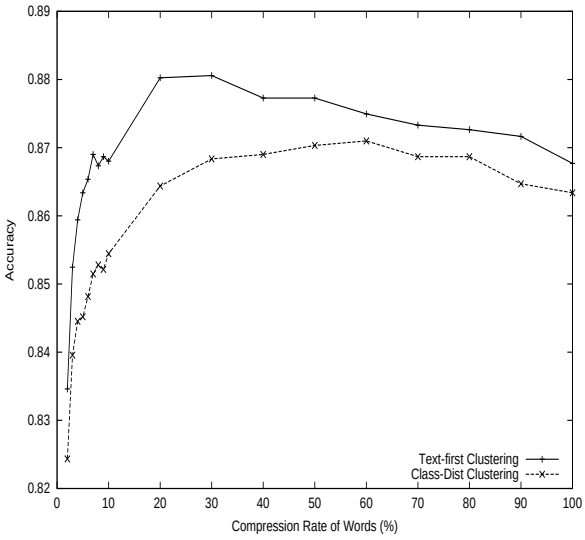

図 2: NB分類器による分類精度

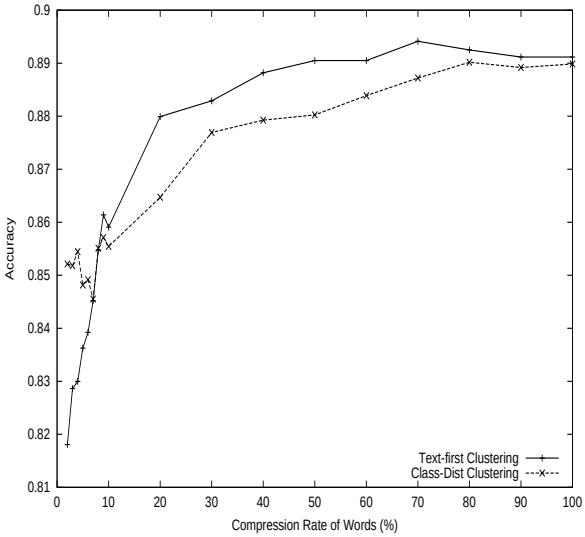

図 3: SVMによる分類精度

表 1: 予測された最適圧縮率

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

表 2: Greedy クラスタリング

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	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

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. この 100%での値の差はそれほど大きくないが, 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わかる. これは, 当初の狙いであった二次元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わかる. しかし, どちらのクラスタリング手法に関しても, 単語圧縮率を下げるに従い精度低下が見られる. これは SVM に対しては単語のクラスタリングを行わない方が良いことを示している. このような性質は, 単語クラスタリングにより精度が向上する NB 分類器と全く逆である.

表 1 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 1 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 2 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, 二次元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, 各カテゴリーが一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, 本稿では一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繋がると思われる.

References

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control*, vol. AC-19, pp. 716–723.

Baker, D. and McCallum, A. 1998. Distributional Clustering of Words for Text Classification. *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pp. 96–103.

Brown, P., Pietra, V.J., deSouza, P.V.,

Lai, J.C. and Mercer, R.L. 1992. Class-based N-gram Models of Natural Language. *Computational Linguistics*, 18(4), pp. 467–479.

Dhillon, I. 2001. Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning. Technical Report 2001-05, UT Austin CS Dept.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning*.

Li, H. and Abe, N. 1998. Word Clustering and Disambiguation Based on Co-occurrence Data. *Proceedings of COLING-ACL 98*, pp. 749–755.

McCallum, A. and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48.

Mitchell, T. 1997. *Machine Learning*, McGraw Hill.

Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3). pp. 103–134.

Rissanen, J. 1987. Stochastic Complexity. *Journal of Royal Statistical Society*, Series B, 49(3), pp. 223–239.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pp. 44–49, Manchester.

Slonim, N. and Tishby, N. 2000. Document Clustering using Word Clusters via the Information Bottleneck Method. *Research and Development in Information Retrieval*, pp. 208–215.

Slonim, N. and Tishby, N. 2001. The Power of Word Clusters for Text Classification. *23rd European Colloquium on Information Retrieval Research*.

Tishby, N., Pereira, F. and Bialek, W. 1999. The Information Bottleneck Method. *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.