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

2003－DBS－130　(20)
2003－ＦＩ－ 71　(20)
2003／5／23

対応分析とベイジアンネットワークを用いた文書分類

福重貴雄　菅野祐司
{fukushige.yoshio, kanno.yuji}@jp.panasonic.com
松下電器産業株式会社
〒140－8632　東京都品川区東品川 4－5－15

文書ベクトルのような高次元データをベイジアンネットワークを用いて分類するには、有効素性の選択による次元削減や適切な離散化が必須の課題となる。筆者らは、単語文書空間における対応分析と MDL 規準に基づいた離散化をベイジアンネットワークに組み合わせて用いることによって、上記の問題の解決を図った。上記方式を二つのベイジアンネット naive Bayes 型と TAN 型と組み合わせて、RWC テキストコーパスを対象として評価実験を行い、Ｆ値で平均 8%（最大 18%）の分類能力の向上を確認した。

Document Categorization using Correspondence Analysis and Bayesian Networks

Yoshio FUKUSHIGE, Yuji KANNO
{fukushige.yoshio, kanno.yuji}@jp.panasonic.com
Matsushita Electric Industrial Co., Ltd.
4-5-15 Higashi-shinagawa, Shinagawa-ku, Tokyo, 140-8632, JAPAN

In utilizing Bayesian networks as a categorizer, it is often problematic when the data to be categorized are represented in a vector form with very high dimension, like document vectors in a vector space model. In this paper, we address this issue by reducing the dimensionality with correspondence analysis (CA) and an MDLP-based discretization, and using the resultant data as the input to a Bayesian network leaner. In our empirical validation with the RWC corpus, this method compares favorably with the conventional results on the same data, showing 8% improvement of F-measure on average (max. 18%)

1. はじめに

文書分類(categorization, classification)は、与えられた文書を、あらかじめ決められた一つ以上のカテゴリのうちの 0 個以上に自動的に分類する技術である。文書分類の手法として、多くの手法が提案されている([15],[9])が、大きく分けて規則ベース型、ベクトル空間型、ニューラルネット型、確率型の手法が挙げられる。本稿で用いるベイジアンネットワークによる文書分類手法は、確率型の手法の一つである。

ベイジアンネットワークは、定性的・記述的な知識と定量的・経験的な知識を組み合わせて利用することに適したツールであり、各種推論、診断において利用され、分類器としても高い潜在能力を持つ。

しかし、ベクトル空間法における文書ベクトルのような高次元データをベイジアンネットワークにおいて利用するには、有効素性の選択による次元削減が重要な課題となる。また、データが実数値を取る場合には、適切な離散化も重要である。

従来のベイジアンネットワークを用いた文書分類においては、素性の選択は、一定の基準を満たす単語のみを選択する（単語フィルタリング）手法がほとんどであった[1]。e.g. Koller and Sahami[11]。しかし、それらは単純な単語頻度に基づくか、単語とカテゴリとの間の相互情報量などに基づく選択方法であった。そのため、頻度の小さい語（しばしば特徴的な語となる）が抜け

[1]単語のクラスタリングによる次元削減を行う例としては、Karčiauskas[9]がある。

落ちてしまったり、(単語とカテゴリの組み合わせ数が膨大であることによる)計算量的な問題を抱えていた。

一方、Deerwester et al. [1]によって、ベクトル空間モデルにおいて、単語文書行列の特異値分解によって、次元数を削減する試み（ＬＳＩ：Latent Semantic Indexing）が提案されている。次元数が削減される効果に加えて、LSI においては、使用される文脈が類似した単語に対しては類似した単語特徴ベクトルが付与されるので、それらを含む文書も類似した文書特徴ベクトルを持つことになり、頻度の低い単語の持つ情報も活用されやすい利点も持つ。この方法は、特定のカテゴリへの分類に対応したものではないので、大域的な次元削減（Sebastiani[15]）の一種である。

しかし、ベイジアンネットワークにおいては確率変数は一般に離散値を取ることが前提となっているため、成分が実数値となる LSI の結果をそのままベイジアンネットワークへの入力とすることはできない。

一般的に、ベイジアンネットワークへの入力のための離散化に関しては、さまざまな方法が提案されているが、Dougherty et al. [3]は、いくつかの手法を対象とした比較実験を行い、Fayyad and Irani[1]による、ＭＤＬ規準を用いた再帰的区間分割による離散化が優れた性能を持つと報告している。また、同手法によって区間分割が認可されなかった軸(素性)に関しては、有効でないとできるので、カテゴリに対応した局所的な次元削減(Sebastiani idid.) の効果も併せ持っている。

筆者らは、単語文書空間の次元圧縮方法として、LSI の代わりに多変量解析の一手法である対応分析を用い、その結果得られる文書特徴ベクトルに Fayyad and Irani による MDL 規準に基づいた離散化を適用することによって、ベイジアンネットワークの上で高次元のベクトルにより表現される大規模な文書データを対象とする手法を提案する。

評価実験として、naive Bayes 型と TAN 型のベイジアンネットワークを用いて、RWC テキストコーパス[9]を対象として文書分類実験を行い、Ｆ値で平均 8%（最大 18%）の分類能力の向上を確認した。

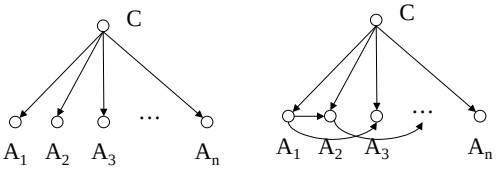
以下、本稿では、2節でベイジアンネットワーク、特に今回の評価実験で用いた naïve Bayes 型ベイジアンネットワークと TAN 型ベイジアンネットワークについて簡単な説明を行い、3節で、大域的な次元削減の手段として採用した主成分分析に関して簡単に説明する。4節では、主成分分析により次元数が削減された文書ベクトルに対して、各カテゴリごとに分類に有効な軸(素性)の抽出と離散化を行う手段として採用した、Fayyad and Irani による、分類のための MDL 規準に基づく連続素性の離散化手法（Fayyad and Irani[1]）に関して説明する。5節では、ベイジアンネットワークにおける条件付確率表の推定において用いたスムージングの手法について説明する。6節で、これらの手法の有効性を確認するために行った RWC コーパスを用いた分類実験について説明し、7節で関連研究について触れ、最後に8節でまとめを述べる。

2. ベイジアンネットワーク

2.1. ベイジアンネットワーク

ベイジアンネットワークは、確率変数間の条件付依存関係を表した非循環型有効グラフ(DAG)で、各種推論などに使われる。各節点は確率変数を表し、(有向)辺は変数間の直接の依存関係を表す。一般には、各変数は離散値をとるとし、各節点には、親節点に対応する変数が与えられたときの、その節点に対応する節点がとる各値について、その値をとる条件付確率を記した条件付確率表(CPT)が与えられている。(根節点の CPT には、事前確率が格納される)

ベイジアンネットワークを用いた分類器では、分類に用いる素性に対応する節点に素性値をセットし、CPT に従った信念伝播などの方法により、カテゴリに対応する節点における信念値を計算し、分類確率とする。



(a) naive Bayes**型**　　(b) TAN**型**

図 1ベイジアンネットワーク

2.2. Naïve Bayes

Naïve Bayes 型のベイジアンネットワークは、木型のベイジアンネットワークのうち、根節点以外の節点がすべて葉であるようなものである。

これは、根節点以外の節点に対応する変数が、根節点に対応する変数の値が与えられた下で条件付独立であることを表している。

今回は、各カテゴリごとに、根節点にカテゴリへの所属を表す変数を対応させ、文書の各素性を表す変数を葉節点に対応させる。このような構成を持つ分類器は naïve Bayes 分類器と呼ばれる。

各素性変数がカテゴリ変数を与えた下で条件付独立であることから、カテゴリ変数に関する事

後確率は

$$p(C|A_1,\cdots,A_n)=\frac{p(A_1,\cdots,A_n|C)\cdot p(C)}{p(A_1,\cdots,A_n)}$$
$$=\frac{p(A_1|C)\cdots p(A_n|C)\cdot p(C)}{p(A_1,\cdots,A_n)}$$

と表される。実際には、$p(A_1,\cdots,A_n)$を求める代わりに、

$$\frac{p(+c|A_1,\cdots,A_n)}{p(\neg c|A_1,\cdots,A_n)}=\frac{p(A_1|+c)\cdots p(A_n|+c)\cdot p(+c)}{p(A_1|\neg c)\cdots p(A_n|\neg c)\cdot p(\neg c)}$$

を計算する。

2.3. TAN(Tree Augmented Naïve Bayes)

TAN(Tree Augmented Naïve Bayes)は、根節点以外の節点間が条件付独立であるという naïve Bayes の制約を少し緩めたもので、根節点以外の節点が、すべて根節点であり、根節点を除いて木構造をなしているようなものを言う。(Friedman and Goldszmidt[5])

TAN 型のベイジアンネットワークを用いた分類器は、素性変数間の依存関係をある程度反映できると同時に、計算量的にも扱いやすい。与えられたデータの下での尤度を最大にするような TAN は、以下のようにして求めることができる。(Friedman and Goldszmidt[5])

① 各素性間の条件付相互情報量 $I_{\hat{P}_D}(A_i,A_j|C)\equiv$

$$\sum_{A_i,A_j,C}\hat{P}_D(A_i,A_j,C)\cdot\log\frac{\hat{P}_D(A_i,A_j|C)}{\hat{P}_D(A_i|C)\cdot\hat{P}_D(A_j|C)}$$

を、求める。

② $A_1,\cdots,A_n$ を節点とするような完全無向グラフを作り、辺に重みとして $I_{\hat{P}_D}(A_i,A_j|C)$ を与える。

③ 上の完全グラフ上での極大生成木(maximum spanning tree)を求め、適当に根節点を選び、辺に向きをつける。

④上の極大生成木の各節点の親として C を加える。

TAN 型のベイジアンネットワークにおいては、

$$\frac{p(+c|A_1,\cdots,A_n)}{p(\neg c|A_1,\cdots,A_n)}=$$
$$\frac{p(A_n|A_{n'},+c)\cdots p(A_2|A_{2'},+c)\cdot p(A_1|+c)\cdot p(+c)}{p(A_n|A_{n'},\neg c)\cdots p(A_2|A_{2'},\neg c)\cdot p(A_1|\neg c)\cdot p(\neg c)}$$

によって、$p(C|A_1,\cdots,A_n)$を求めることができる。

ただし、A_1 の親は C のみとし、$A_{i'}$ は、A_i の親となっている素性$(i=2,\cdots,n)$とする。

3. 対応分析を用いた大域的次元削減

3.1. LSI(Latent Semantic Indexing)法

LSI 法は、Deerwester[1]により提唱された、特異値分解による単語文書行列の低次元近似により、次元削減を行う手法である。

単語文書行列を $F=UDV^T$ と特異値分解し、大きい順に k 個までの固有値に対応する部分を取り出し、$\tilde{F}=U_kD_kV_k^T$ により F を近似し、文書ベクトルを $\hat{\mathbf{f}}_i=V_i^T\mathbf{f}_i$ により近似する。

3.2. 対応分析

対応分析(correspondence analysis)は、二つの離散変数間の関係を分析する多変量解析の一手法で、質的データに対する主成分分析的な側面を持つ。

主成分分析が、データ行列（＝単語文書行列）$F=(f_{i,j})$をそのまま使って特異値分解を行うのに対して、対応分析においては、F の成分の行方向の和を対角要素に持つ行列 $G\equiv diag(g_i)=diag\left(\sum_j f_{i,j}\right)$ および

F の成分の列方向の和を対角要素に持つ行列 $H\equiv diag(h_j)=diag\left(\sum_i f_{i,j}\right)$ として、

$G^{-\frac{1}{2}}FH^{-\frac{1}{2}}=\left(\frac{f_{i,j}}{\sqrt{g_i}\sqrt{h_j}}\right)$ を $=UDV^T$ と特異値分解し、$\mathbf{x}_i=\left(\frac{u_{i,2}}{\sqrt{g_i}}\ \cdots\ \frac{u_{i,k+1}}{\sqrt{g_i}}\right)$ を求め、さらに長さ 1 に正規化したものを、文書 i の文書特徴ベクトルとした。上記のように、対応分析においては、周辺頻度の-1/2 乗による修正が入るので、頻度の小さい語や簡潔な文書の影響が強くなるので、LSI 法や主成分分析に比べて、分類タスクに適した圧縮結果が得られると予想される。

4. ＭＤＬ規準による離散化と局所的次元削減

MDL 規準は、Rissanen[17]により提唱されたモデル選択の基準であり、モデル自身を記述するための最小記述長とそのモデルの下でのデータを記述するために必要な最小記述長の和(=系の MDL)が最小になるようなモデルを選択する、というものである。

Fayyad and Irani[1]は、連続値データの集合を区間に分割する方法として、分割による情報利得が最大になる点における二分割を、分割による系の MDL が減少しなくなるまで再起的に繰り

返すという方法を提案している。

具体的には、

式 1: $Gain(A,T;S) > \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T;S)}{N}$

が成り立つときに限り、分割点 T で、区間 S を区間 S_1 と S_2 に分割することを許す。ただし、

$$Gain(A,T;S) \equiv Ent(S) - \frac{|S_1|}{N} Ent(S_1) - \frac{|S_2|}{N} Ent(S_2),$$

$$\Delta(A,T;S) \equiv \log_2(3^k - 2) - [kEnt(S) - k_1 Ent(S_1) - k_2 Ent(S_2)],$$

$$Ent(S) \equiv -\sum_{i=1}^{k} P(C_i,S)\log_2(P(C_i,S))$$

で、$P(C_i,S)$は、区間Sに入る学習データのうち、カテゴリ C_i に属すものの割合であり、k は、区間Sに入る学習データのクラスの種類数である。($Ent(S_1)$, $Ent(S_2)$, k_1, k_2 についても同様)

$Gain(A,T;S)$は、分割による情報利得である。

筆者らは、各カテゴリごとに、この MDL 規準による区間分割を、対応分析の結果得られた各軸(主軸)に対して行い、離散化のための区切り点を得、各文書特徴ベクトルの成分に対して、対応する軸のどの区間に入るかを示す区間番号を離散化結果として、ベイジアンネットワークへの入力で用いる素性の値とした。

同時に、最初の段階で式1を満たさず、分割できない軸は、そのカテゴリに関する分類には有効でないとして除いて、ベイジアンネットワークへの入力とする素性には含めなかった。これにより、そのカテゴリへの分類に必要な軸を絞り込む局所的な次元削減を行っている。

5. スムージング

ベイジアンネットワークの各節点には、親節点の値(の組)が与えられたときに、その節点がとりうる各値に対する条件付確率を格納した条件つき確率表(CPT)が付与されている。

CPT の内容は、訓練用データから推定されるが、親節点（の組）が特定の値をとるようなデー数が少ない場合は、推定される条件付確率も、本来の値からかけ離れてしまう可能性が大きい。

このような標本数の少なさによる推定誤差の増大を避けるために、スムージングを行った。

よく行われているスムージングの例は、ベルヌーイ試行の成功確率θを推定する際に、試行回数を n、そのうち成功数が y であったときに、θ の推定値 $\hat{\theta}$ として、

$\hat{\theta} = \frac{y}{n}$ の代わりに、$\hat{\theta} = \frac{y+1}{n+2}$ または、$\hat{\theta} = \frac{y+0.5}{n+1}$ とする例である。これは、実際には、θ の分布型としてβ分布を仮定し、事前分布として $Beta(1,1)$ または $Beta(0.5,0.5)$ を仮定していることに相当する。

今回は、（親節点の値 $\boldsymbol{\pi}_X$ が与えられた下で）各素性 X がパラメータ $\boldsymbol{\Theta}_X$ を持つ多項分布に従い、また、$\boldsymbol{\Theta}_X$ がパラメータ

$$\left(N^0_{X|\Pi_X} \cdot \frac{N(X=1)}{N} \quad \cdots \quad N^0_{X|\Pi_X} \cdot \frac{N(X=n)}{N} \right)$$

を持つ Dirichlet 分布に従うと仮定し、親節点の値 $\boldsymbol{\pi}_X$ が与えられた下で素性Xが値xを取る確率の、データ集合 D に基づく推定値を

$$\frac{N(\boldsymbol{\Pi}_x = \boldsymbol{\pi}_X, X = x) + N^0_{X|\Pi_X} \cdot \frac{N(X=x)}{N}}{N(\boldsymbol{\Pi}_x = \boldsymbol{\pi}_X) + N^0_{X|\Pi_X}}$$

により推定する。ただし、N は、訓練データ数、$N(P)$ は、命題 P が成り立つ訓練データ数とする。この $N^0_{X|\Pi_x}$ は、事前分布の強さを表すもので、「事前標本数」と呼ばれることもある（たとえば Gelman et al.[7]）

6. 評価実験

6.1. 実験データ

評価実験には、ＲＷＣテキストコーパス第 2 版の、毎日新聞記事 UDC(国際十進分類法)コード付与データ(RWC-DB-TEXT-95-3)を用いた[9]。

同データは、毎日新聞 1994 年の約 3 万件の記事に人手で UDC コード（[1]）を付与したものである。

そのうち、平・春野[20]において用いられた、学習用 1,000 記事、テスト用 1,000 記事を実験用データとして使用した。表 1に、実験データのカテゴリ別のデータ数を示す。

表 1実験データのカテゴリ別内訳

カテゴリ名	訓練データ数	テストデータ数
スポーツ	161	147
犯罪(刑法)	155*	148
政府	135	142
教育システム	110	124
交通	113*	103
軍事	110	118
国際関連	96	97
言語活動	76	83
演劇	86	95
作物	72	78

*)平・春野[20]では、記事 940105242 が、刑法（犯罪）に入っていたが、実際のコードは交通であったので、修正した。

6.2. 実験手順

前処理として、以下のようにして、上記データから単語の切出しを行っておく。

① 筆者らが有する単語辞書（約 43 万語）を用いて、極大切り出し（他の単独の見出しに被覆されていれば切り出さないとする単語切り出し方法,[12]）により、上記データを含む RWC コーパスの UDC が付与されている約 3 万記事全体について、単語切り出しを行う。

② 切り出し結果から、約 3 万語の不要語辞書を参照し、不要語を除く。

この切出し結果を用いて、以下の(1)～(3)の実験を行った。なお、実験においては、単語切り出し、対応分析（LSI）は、WS 上で行い、以降は PC 上の R システム[25]を使った。

(1) 次元削減方法、正規化、ベイジアンネットの型を変えた実験

次元削減方法、正規化の有無、ベイジアンネットの型を表 2のように変えて評価実験を行った。

表 2 実験設定

実験コード	次元削減	正規化	ベイジアンネットの型
CA+TAN	CA	あり	TAN
CA+NB	CA	あり	naive Bayes
CA+TAN(RAW)	CA	なし	TAN
LSI+TAN	LSI	あり	TAN

単語文書行列の重みは、対応分析を行うときは、単純頻度を、LSI を行うときは tf.idf を用いた。

なお、スムージングの「事前標本数」 $N^0_{\Pi_x}$ については、素性／カテゴリによらず 10 とした。

(2) 訓練デー多数を変えた実験

訓練データ数を 750, 500, 200, 100, 75 とした比較も行った。ここでは、各データ数別に 10 回ずつ、訓練用の 1,000 記事からランダムに訓練用の記事セットの抽出を行い、それを用いて学習を行った結果を、テスト用記事 1,000 で評価し、F 値の平均をとった。次元削減は CA、正規化あり、で、naive Bayes と TAN のそれぞれについて評価を行った。事前標本数は 10 とした。

(3) 事前標本数を変えた実験

訓練用の 1,000 記事を用いた実験で、事前標本数を 0,1,10,20,30 と変えて実験を行った。次元削減は CA、正規化ありとした。

6.3. 評価指標

評価指標としては、F 値([20])を用いた。ここで、F 値は、次のように定義される 0～1 の値をとる数であり、大きいほど分類器の性能がよいとされる。

$$F_\beta \equiv \frac{(\beta^2+1)\cdot prec\cdot rec}{\beta^2\cdot prec + rec}$$ ただし、

$$prec = \textbf{適合率} = \frac{TP}{TP+FP}, rec = \textbf{再現率} = \frac{TP}{TP+FN}$$

であり、TP,FP,FN,TN は、以下のような分類器の判定と本来の所属パタンを持つテストデータの数である。

	分類器の判定	本来の所属
TP	所属	所属
FP	所属	非所属
FN	非所属	所属
TN	非所属	非所属

β は、適合率と再現率の評価重みを制御するパラメータで、0 なら F 値は適合率に、∞なら再現率に一致する。今回は β =1 とした。

6.4. 結果

次元削減法・正規化・ベイジアンネットワークの型を変えた比較実験結果を表 3に示す。

これらを見ると、対応分析による次元削減結果を正規化した場合に、LSI 法や、正規化しない場合に比べて優れた結果が得られている。

また、平・春野[9]で報告されているトランスダクティブ・ブースティング法(表中 TB。数値は平・春野[9]中の数値)による結果と比べても良好な結果が得られている。ただし、平・春野[9]で用いられているデータは、全体としては同じであるが、訓練データとテストデータの切り分け方が異なるので、直接の比較はできない。

学習データ数を変えた場合の実験結果を表 4、図 2 に示す。TAN による分類は、学習データ数が 200 を超えた領域では naive Bayes より優れた結果を示しているが、それ以下の領域では、naive Bayes の方が優れた結果を示している。これは、TAN 型のネットワークは naive Bayes より複雑であるため、より多くのデータを必要とするということが現れていると考えられる。平・春野[9]での実験結果との比較すると、今回の手法が学習データ数 200 以上の領域で良好な結果を示している。

データ数が 75 および 100 の場合に、急激な指標の低下が見られる。とくに表 4において、*印をつけた設定では、試行した中にカテゴリに所属すると判定された文書が全くない場合があった（そのような場合、F 値は 0 とした）。

そのような場合を見てみると、有効軸数が極端に少なくなっている（表 6）。

表 3　訓練データ数 1000 での F 値

カテゴリ名	CA +TAN	CA +NB	CA +TAN (RAW)	LSI +TAN	TB
スポーツ	0.93	0.92	0.88	0.88	0.90
刑法	0.80	0.80	0.80	0.70	0.75
政府	0.77	0.76	0.58	0.69	0.72
教育	0.86	0.89	0.89	0.75	0.78
交通	0.79	0.77	0.78	0.63	0.70
軍事	0.78	0.75	0.72	0.66	0.78
国際関連	0.66	0.65	0.58	0.58	0.56
言語活動	0.78	0.74	0.82	0.74	0.69
演劇	0.89	0.88	0.86	0.81	0.86
作物	0.93	0.90	0.90	0.86	0.85
平均	0.82	0.81	0.78	0.73	0.76

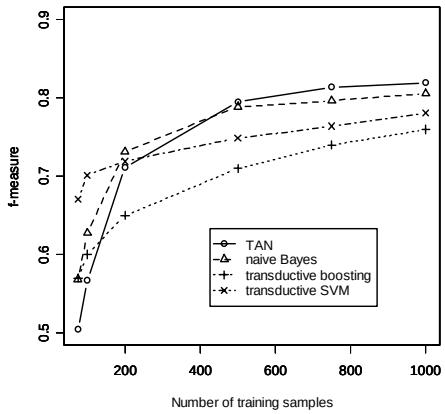


図 2　訓練データ数と F 値の平均

表 4　訓練データ数と F 値

(a) Naïve Bayes

カテゴリ名	学習データ数					
	75	100	200	500	750	1000
スポーツ	0.66	0.70	0.80	0.89	0.90	0.92
刑法	0.53	0.62	0.74	0.76	0.79	0.80
政府	0.58	0.62	0.67	0.74	0.74	0.76
教育	0.54	0.73	0.83	0.87	0.88	0.89
交通	0.47*	0.52	0.63	0.75	0.77	0.77
軍事	0.55	0.51*	0.66	0.74	0.74	0.75
国際関連	0.37*	0.49	0.58	0.62	0.64	0.65
言語活動	0.57	0.58	0.68	0.75	0.74	0.74
演劇	0.62	0.64*	0.81	0.87	0.87	0.88
作物	0.81	0.86	0.89	0.91	0.89	0.90
平均	0.57	0.63	0.73	0.79	0.80	0.81

(b) TAN

カテゴリ名	学習データ数					
	75	100	200	500	750	1000
スポーツ	0.64	0.68	0.79	0.90	0.92	0.93
刑法	0.49	0.56	0.73	0.76	0.79	0.80
政府	0.53	0.60	0.67	0.74	0.77	0.77
教育	0.45	0.65	0.80	0.87	0.87	0.86
交通	0.42*	0.50	0.62	0.76	0.78	0.79
軍事	0.45*	0.48*	0.67	0.76	0.78	0.78
国際関連	0.33	0.35*	0.53	0.62	0.65	0.66
言語活動	0.47	0.48	0.64	0.76	0.77	0.78
演劇	0.55	0.60*	0.80	0.86	0.87	0.89
作物	0.71	0.76	0.87	0.93	0.93	0.93
平均	0.50	0.57	0.71	0.80	0.81	0.82

表 5　訓練データ数と有効軸数

カテゴリ名	学習データ数					
	75	100	200	500	750	1000
スポーツ	15.9	24.5	39.0	96.0	140.8	167
刑法	17.7	17.2	32.0	53.8	81.8	111
政府	15.4	24.3	31.9	76.4	111.4	132
教育	20.9	27.9	35.6	78.6	103.9	127
交通	16.3	23.6	29.5	54.4	74.3	84
軍事	14.0	16.9	34.3	91.0	139.3	181
国際関連	13.7	15.4	21.7	45.9	63.1	87
言語活動	25.0	29.8	42.8	54.4	182.6	215
演劇	23.5	23.7	38.7	101.9	152.1	191
作物	30.6	44.9	71.6	138.0	185.8	219
平均	19.3	24.8	37.7	88.5	123.5	151.4

表 6　F 値=0 となった場合

カテゴリ名	訓練データ総数	訓練所属文書数	有効軸数	ベイジアンネットの型
軍事	75	5	3	TAN
交通	75	5	2	TAN/NB
国際関係	75	3	10	NB
国際関係	100	7	4	TAN
軍事	100	7	2	TAN/NB
演劇	100	4	1	TAN

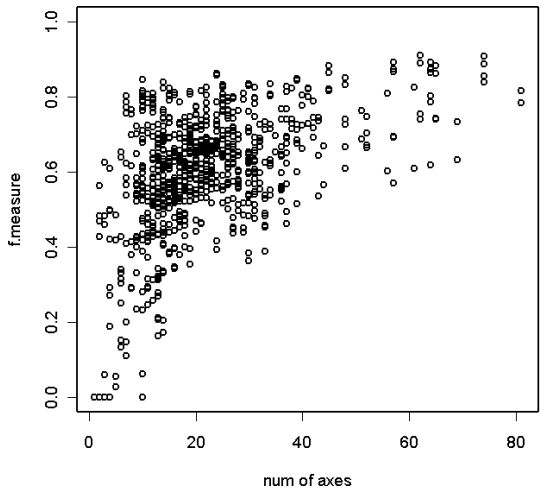


図 3　有効軸数とF値（訓練数 75, 100）

したがって、MDL 規準による分割条件(式 1)を少し緩めることが必要かもしれない。ただし、訓練データ数 75, 100 における有効軸数と F 値をプロットした図 3 をみると、少ない軸数でも分類性能が高い場合もあるので、一概に軸が少ないことが悪いとも言えないので、詳細な分析が必要である。

表 7　事前標本総数と F 値の変化

カテゴリ名	事前標本総数				
	0	1	10	20	30
スポーツ	0.13	0.94	0.94	0.93	0.92
刑法	0.33	0.78	0.80	0.81	0.80
政府	0.19	0.74	0.76	0.79	0.77
教育	0.05	0.88	0.86	0.87	0.85
交通	0.38	0.80	0.80	0.78	0.77
軍事	0.11	0.79	0.79	0.79	0.79
国際関連	0.54	0.64	0.66	0.66	0.66
言語活動	0.13	0.73	0.80	0.78	0.77
演劇	0.08	0.89	0.88	0.87	0.86
作物	0	0.90	0.91	0.93	0.94
平均	0.19	0.81	0.82	0.82	0.81

事前標本数を変化させた実験においては、事前標本数を 1 から 30 まで変化させても、平均的にはそれほど違いは見られなかった。しかし、個々に見てみると、いくつかのカテゴリにおいては、事前標本数の差によって、F 値にかなりの差が出ているものもある。スムージングをまったく行わなかった場合は、極端に悪い結果となっている。これは、尤度比の計算は積となっており、一部でも低いものがあれば全体が低くなるという仕組みが影響を与えていると考えられる。

7. 関連研究

秋葉[1]は、ベイジアンネットを使った自然言語処理に関する紹介である。

次元削減に対応分析を用いた例としては、Payne and Edwards[15]がある。Payene らは、削減結果を用いて、ユークリッド距離を用いた最近隣法(nearest neighbor method)による分類実験を行っている。

RWC テキストコーパスを用いた文書分類の研究としては、山崎・イド[27], 平・向内・春野[23], 平・春野[20],平・春野[9], 桂田[10],がある。

8. まとめと今後の課題

ベクトル空間法における文書ベクトルのような高次元データをベイジアンネットワークにおいて利用するには、有効素性の選択による次元削減が重要な課題となる。また、データが実数値を取る場合には、適切な離散化も重要である。

筆者らは、単語文書空間の次元圧縮方法として、多変量解析の一手法である対応分析を用い、その結果得られる文書特徴ベクトルに MDL 規準に基づいた離散化を適用することによって、ベイジアンネットワークの上で高次元のベクトルにより表現される大規模な文書データを対象とする手法を提案した。

評価実験として、naive Bayes 型と TAN 型のベイジアンネットワークを用いて、RWC テキストコーパスを用いた文書分類実験を行い、提案手法が、これまで同データに対して得られている、SVM、トランスダクティブブースティング、トランスダクティブ SVM による結果と比較して、F 値で平均 8%（最大 18%）の分類能力の向上を確認した。

ベイジアンネットワークの特長は、定性的・記述的な知識と定量的・経験的な知識を組み合わせて利用することに適した枠組みを持っていることである。

ただし、今回の手法は、そうした特徴を活かしていない。今後の課題として、人間が持っている定性的な知識を取り入れて分類性能を向上させることが挙げられる。

また、特にデータが少ない場合に系の複雑さを調整することも必要である。一つの方法として、Sahami[17]における KDB アルゴリズムを検討したい。

離散化においても、本稿の手法では、ネットワークの構造を意識しないで離散化を行う Friedman and Goldszmidt[6]のような、ネット

ワークの構造に応じた離散化も検討したい。

また、本稿の手法では、スムージングは行っているものの、基本的に推定は点推定であり、Bayesian 的な、推定量の確率分布を考えていない。ブースティングなどの協調学習やモデル平均化とのつながりも含めて検討していきたい。

謝辞

毎日新聞 94 年版の使用に関して、記事データの研究利用を許諾してくださった毎日新聞社に感謝いたします。

また、比較対照のために実験文書セットを公開してくださったＮＴＴコミュニケーション科学基礎研究所の平博順氏に感謝いたします。

参考文献

[1] 秋葉友良: 自然言語処理におけるベイジアンネット，人工知能学会誌，Vol.17, No.5, pp.553-558, 2002.

[2] Deerwester, S., Dumais, S. T., Furnas, G.W., Landauer, T.K., and Harshman, R.: Indexing by latent semantic indexing. Journal of the American Society for Information Science Vol. 41, No.6, pp.391-407, 1990.

[3] Dougherty, J., Kohavi, R. and Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features, Proceedings of the Twelfth International Conference on Machine Learning, pp.194-202, 1995.

[4] Fayyad, U.M. and Irani, K. B.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022 - 1027, 1993.

[5] Friedman, N., Geiger, D. and Goldszmidt, M.: Bayesian Network Classifiers, Machine Learning, Vol.29, pp.131-161, 1997.

[6] Friedman, N. and Goldszmidt, M.: Discretizing Continuous Attributes While Learning Bayesian Networks, Proceedings of the 13th International Conference on Machine Learning, pp. 157-165, 1996.

[7] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B:Bayesian Data Analysis, Chapman & Hall/CRC, 1995.

[8] (社)情報科学技術協会: 国際十進法分類日本語中間版第 3 版. 丸善,1994.

[9] Karčiauskas, G.: Text Categorization Using Hierarchical Bayesian Network Classifiers, M.Sc. thesis. Aalborg University, 2002.

[10] 桂田浩一,小山誠,大原剛三,馬場口登,北橋忠宏: 文書分類システムの誤りに着目した分類ルール修正法,情報処理学会論文誌,Vol 43, No.6, pp.1880-1889, 2002.

[11] Koller, D. and Sahami, M.: Toward Optimal Feature Selection, International Conference on Machine Learning, pp.284-292, 1996.

[12] 倉知一晃, 野口直彦, 菅野祐司, 稲葉光昭: 日本語文書に対する新しい索引検索方式--索引作成と今朝区の原理--, 第 50 回情処全大, 4F-2, 1995.

[13] 永田昌明,平博順：テキスト分類－学習理論の「見本市」, 情報処理, Vol.42, No.1, pp.32-37, 2001.

[14] 大津起夫:社会調査データからの推論:実践的入門. 甘利俊一他編　言語と心理の統計,岩波書店, pp.129 - 177, 2003.

[15] Payne, T.R. and Edwards, P.: Dimensionality Reduction through Correspondence Analysis, AUCS/TR9910, University of Aberdeen, Scotland, 1999.

[16] Pearl, J.: Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann., 1988.

[17] Rissanen, J.: Modeling by shortest data description, Automatica, Vol.14, pp. 465-471, 1978.

[18] Sahami, M.: Learning Limited Dependence Bayesian Classifiers, Proceedings of the Second International Conference of Knowledge Discovery and Data Mining, pp. 335-338, 1996.

[19] Sebastiani, F.: Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No.1, pp.1-47, 2002.

[20] Sundheim, B.M.: Overview of the Fourth Message Understanding Conference. Proceedings of Fourth Message Understanding Conference, pp.3-29, 1992.

[21] 平博順, 春野雅彦: Support Victor Machine によるテキスト分類における属性選択, 情報処理学会論文誌, Vol. 41, No.4, pp.1113-1123, 2000.

[22] 平博順,春野雅彦: トランスダクティブ・ブースティング法によるテキスト分類. 情報処理学会論文誌 Vol. 43 No.6, pp.1843－1851, 2002.

[23] 平博順,向内隆文,春野雅彦: Support VectorMachine によるテキスト分類, 情報処理学会研究報告 NL-128-24, pp.173-180, 1998.

[24] 竹内広宣, 小林メイ, 青野雅樹, 寒川光: 多変量解析に基づいた情報検索手法の比較検討. 情報処理学会研究報告, Vol. FI 66-12, pp.87-93, 2002.

[25] The R Project for Statistical Computing (http://www.r-project.org)

[26] 豊浦潤, 徳永健伸, 井佐原均, 岡隆一: ＲＷＣコーパスにおける分類コードつきテキストデータベースの開発. 情報処理学会研究報告,Vol. NL 114-5., pp. 27-32, 1996.

[27] 山崎毅文, イドダガン: 誤り駆動型学習とシソーラスを用いた文書自動分類, 情報処理学会研究報告 NL-120-14, pp.89-96, 1997.

[28] Yang, Y. and Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, Proceedings of the 14th International Conference on Machine Learning, pp. 412-420, 1997.