DEWS2004 3-B-03

分類階層構造を考慮した文書の自動分類

柳田　卓郎†　　三浦　孝夫†

† 法政大学 工学研究科 電気工学専攻　〒 184–8584 東京都小 金井市梶野町 3-7-2
E-mail: †{i02r3244,miurat}@k.hosei.ac.jp

あらまし　本研究では分類階層構造を考慮した文書データの自己組織化マップ（以下 SOM) [4] による自動分類について論じる。SOM によるデータ分類手法のなかで、分類器でありながらトポロジを保持することができるものとして *TaxSOM* [1] が知られている。TaxSOM では教師なし学習によって全データから効果的な訓練事例となりうるものを抽出しながら学習を進めるため、明確に試験データと訓練データを分別する必要がなくなり、分類のために必要となるデータの総数を軽減することができる。本稿では出力層上で横の学習を追加することにより、隣接するユニット間の関連を強くし、更なる学習効果の一定の向上が可能となることを示す。本稿では k 次伝播学習モデル（以下 *TaxSOM(k)*) を提案する。分類精度を再現率と適合率によって評価することで TaxSOM(k) の分類器としての有効性を検証する。
キーワード　分類階層構造，マルチクラス, 自己組織化マップ, 機械学習, 分類，TaxSOM(k)

Classifying Documents according to Taxonomy

Taqlow YANAGIDA† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †{i02r3244,miurat}@k.hosei.ac.jp

Abstract　This paper proposes classifying documents by Self-Organizing Maps according to Taxonomy. The proposed model, named TaxSOM(k), is based on revision of Self-Organizing Maps which involves learning mechanism to neighbors. Our goal is to obtain good classifier for multi-class dataset. We show some experimental results, to see TaxSOM(k) maps are able to illustrate the output of clustering and classification at the same time.
Key words　Taxonomy, Self-Organizing Maps, multi-class, Machine learning, classification, TaxSOM(k).

1. 前　書　き

近年、ニュース記事などの文書データのようなカテゴリ階層を持つ情報の有効活用に対する研究が活発に行われている。ドキュメントは少なくともひとつ以上のカテゴリに分類されるマルチクラスデータである。各ラベルは”A”のような記号的なものではなく、”Economy”のような言語的意味を持ったものであらわされる。ラベルが文書内容の抽象度程度をしめすことで, カテゴリがもつ階層とラベルを対応付ける。本稿ではデータ間の関係を明確に示すために、分類階層構造を考慮した自動分類手法をしめす。文書データの自動分類で高い精度を得られる手法として、*Support Vector Machine(SVM)* [3] が知られている。SVM は教師あり学習によるバイナリ分類を行う。教師あり学習はあらかじめ与えられた訓練事例を元に学習を行い、分類ルールを構築する手法である。この手法の問題点は分類対象全てのパターンを網羅した大量のデータを分類したいデータとは別に事前に用意する必要がある点である。しかし、マルチクラスを持つ文書データを分類するうえで、このような大量の訓練事例を用意することは非常に困難である。本稿では、与えられたデータから学習に効果的に働くデータを訓練事例として抽出する TaxSOM を利用する。これにより試験データと訓練データを明確に分別する必要をなくし、分類のために必要となるデータの総数を軽減する。本稿では TaxSOM の学習効果をさらに高めるために、学習の k 次伝播機構を導入した TaxSOM(k) モデルを提案する。2 章では関連研究として SOM と TaxSOM について述べる．3 章で提案手法である TaxSOM(k) について論じ，4 章では実験と評価方法について考察する。5 章では実験結果を示し，6 章で結びとする．

2. 関 連 研 究

2.1 文書データの分類

テキスト分類は文書を事前に与えられた複数のカテゴリに振り分けるための技術である [11]。具体的な例としては, 新聞記事を記事の内容から政治, 経済, 娯楽といったカテゴリに自動的に振り分けるために用いられる。自動分類では文書に対する分類ルールを機械学習によって導く必要がある。学習する文書デー

タは、各単語の出現頻度を用いた高次元ベクトルを用いて数値変換されるのが一般的だが, 高次元すぎるデータには過学習に対する懸念がつきまとうため次元の縮小が必要となる。現在、SVM が高次元ベクトルによる過学習の問題を解決し, 最も高い分類精度を持つ分類器であることが検証されている。しかし、SVM はバイナリ分類であるため、複数クラスに属するデータの分類には適さないと考えられる。このことは Luis Gravano らが行ったマルチクラス Web データの SVM による分類の結果から検証されている [5]。

2.2 自己組織化マップ (SOM)

SOM は教師なし競合学習によってデータ同士の近さによるクラスタ化を行い、その結果を 2 次元格子マップ (以下 SOM マップ) に可視化することができる。SOM を分類器とみなすとき、実験者は各点のラベルを分類結果と考えることになるが、ラベルの決定は勝者全奪であるためクラスをまたぐ特徴をもったデータに対する分類結果を反映することができない。例えばある点にクラス 1 のデータが 3 件, クラス 2 のデータが 1 件割り当てられたとき, 点にはラベル 1 が与えられる. 本研究では訓練データによる学習を行った後, 試験データを分類させる教師あり学習により実験を行う. この狙いは試験データを分類させることで各点の正答率を表すことにある. マップ上の各格子点は位置情報とコードブックベクトル $\vec{w}_i = [w_1, ..., w_m]$ を持つ。値 w_m は各点のコードブックの要素である。SOM マップは入力データによる競合フェーズと協調フェーズから生成される。競合フェーズで入力ベクトル $\vec{x}$ に最も近いコードブックが勝者ユニットとして選択される。協調フェーズでは勝者ユニットの周囲のコードブックが入力ベクトルに近くなるように次式によって修正される。

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \alpha(t) * h_{i,i'}[\vec{x} - \vec{w}_i(t)] \quad (1)$$

$\alpha(t)$ は学習率係数であり、$h_{i,i'}$ は点 i と勝者点 i' 間の距離を考慮した近傍関数を示す。我々は SOM マップから大まかなクラスタを読み取ることができる。しかし、SOM の学習は入力層から計算層を経て出力層に至るまでの間に縦一本のつながりしか存在しない。そのため出力層上のユニットは隣り合っていても学習時の関連は薄いという問題がある。また、学習によって得られた SOM マップは客観的評価方法を持たず、結果が非常に曖昧なものとなってしまう。さらに、我々は、文書データの分類では SOM がトポロジを保持しないことをこれまでの研究 [7] で示している。文書データの学習から得た SOM マップの例を図 1 に示す。

2.3 TaxSOM

これまで Adami らは TaxSOM [1] を提案し、教師なし学習法であった自己組織化マップ (SOM) の分類にトポロジが保持できることを示した。TaxSOM 学習は、競合フェーズにおいて学習に効果的に働くデータを全データによるバッチ学習のなかで抽出し、動的に教師データを生成する。SOM の協調フェーズにおいて各格子点に振り分けられた文書ベクトルの重心 $\vec{x}_j(t)$ から

$$\vec{w}_i(t+1) = \frac{\sum_j n_j h_{i,j} \vec{x}_j(t)}{\sum_j n_j h_{i,j}} \quad (2)$$

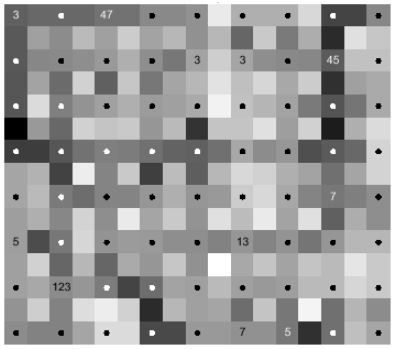

図 1　SOM マップの例

によってコードブックを算出する。n_j はボロノイ集合 j 内に含まれるデータの総数である。さらに、コードブックには各点のラベルが次式

$$\vec{w}'_i(t+1) = f(\vec{w}_i(t+1), L_N(i)) \quad (3)$$

によってエンコードされる。$L_N(i)$ は点 i 上のラベル集合である。$L_N(i)$ は実験データの持つクラスに合わせて 7 次元のバイナリベクトル $\vec{C}_j = [C_1, C_2, ..., C_7]$ で表し，競合フェーズで各点に配置されたデータの持つクラス c を基に値 C_c を 1 とすることで，実験者が与えるものとする。配置されたデータの持つクラス c が、各点の $L_N(i)$ に含まれるとき値 C_c を 1 としてエンコードする。この改善によって従来の SOM では得られなかった、分類器でありながらトポロジを保持するマップを生成可能にする。TaxSOM の狙いは、正確な分類に寄与するデータに対してラベル情報を付加することで、ベクトル間の位置関係をより明確に示すことである。TaxSOM 学習から得られたマップを図 2 に示す。従来の SOM の結果 (図 1) と比較すると同じデータで学習を行った TaxSOM マップ (図 2) のほうがクラスタをよりはっきり確認できる。また、各点のラベルから同じクラスがマップ上でより近い距離に配置されており、トポロジが保持されていることも確認できる。

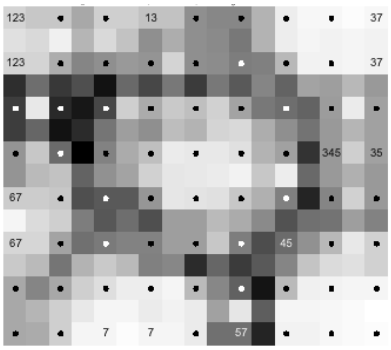

図 2　TaxSOM マップの例

3. k 次伝播モデルによる TaxSOM の拡張

TaxSOM による学習から我々はトポロジを保つ分類器を得る

ことができる。しかし、TaxSOM を分類器として考える場合、その能力はあまり高くないことが Adami らの研究 [2] で検証されている。また、TaxSOM の学習部分は従来の SOM と同様に入力データから計算層の 1 ユニットを経て出力層の 1 ユニットまで縦一本のつながりしか持っていないため出力層上のユニットは隣り合っていても学習時の関連は SOM と同様に薄いままである。

既に著者らは、SOM 学習を拡張した k 次伝播 SOM(以下 SOM(k)) による文書分類 [8] を行っている。我々は SOM(k) から、良好な分類精度を得ることはできたが、トポロジを保つことができなかった。その結果、マップによるクラスタリングとデータの分類との関係が希薄になってしまい、SOM マップ本来の良さを利用することができなかった。

そこで我々は TaxSOM の出力層上で従来の勝者ユニットとその近傍に対する学習に加え、近傍の外側への横方向の k 次伝播学習を追加することで隣接するユニット間の関連を強くし、トポロジを保持しながら更なる学習効果の向上を図る。この拡張によってマップ上のクラスタ化出力とクラス分類出力を一致させることを考える。このモデルを我々は TaxSOM(k) と呼ぶ。TaxSOM(k) では競合フェーズで得られた勝者点から近傍に対して学習パラメータとクラス情報を伝播する。実験者が伝播数 k を決定するものとし、伝播パラメータが影響を及ぼす範囲を制限できる。各点は伝播されたクラス情報を分布として保持することで複数のクラスをまたぐ特徴を持ったマルチクラスデータの分類を可能にする。自然界の伝播法則を参考にすると、力学モデル [6] と波動モデルを考えることができる。力学モデルはある一点への周囲からの複数の入力ベクトルは 1 つの伝播ベクトルに合成されるのに対して、波動モデルでは複数の入力ベクトルから複数の干渉波が生成され、多方向への伝播が起こる. 勝者点は, 特徴のある傾向を持ったベクトルを持っており, マップ上に現れるクラスタの基点となっていることが多い。このことから 1 点対 1 点の伝播しかなされない力学モデルよりも, 勝者点を中心とした周囲のユニットとの関係を学習する波動伝播モデルの方がクラスタの強調には有効であると考えられる。したがって本稿では伝播方式に波動モデルを適用する。

波動モデルによる伝播を考慮したコードブックベクトルは, 次式

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \sum_j \left\{ \exp(\frac{\beta}{k_j t_j}) * \frac{A_j}{\sqrt{r_j}} \cos(\omega_j t_j) \vec{w}_j(t_j) \right\} \quad (4)$$

で算出する。ω は角速度を示し、ユニット j を基点とする伝播数 k_j とともに減衰する。r_j は勝者点と伝播点との距離であり、この距離が大きくなるほど波の振幅 A_j は減衰する。コードブックへのラベルのエンコードは、TaxSOM と同様にする。この手法から、従来のマップにくわえて、クラス分布情報を結果として獲得することができる。TaxSOM マップ上の各点のクラス分布情報を表す出力を LVQ マップと呼ぶ。LVQ マップから、各点に配置されたデータの傾向を判定する。図 1,2 で使ったデータから学習によって生成した図 2 の LVQ マップの例を表 1 に示す。表の左端の"Point"は TaxSOM マップ上の点 (x,y) を示す。"1"～ "7"は, 学習時に点 (x,y) にクラス 1～7 のベクトルが出現した頻度を示している。右端の"Total"は前述した各点のベクトル出現頻度の総和である。

Point	1	2	3	4	5	6	7	Total
(2,1)	1	2	2	.00	.00	.00	.00	5.0
(2,2)	.00	.00	.00	.00	2	.00	.00	2.0
(2,3)	.00	8	.00	.00	.00	.00	.00	8.0
(3,0)	.00	.00	.00	.00	.00	3	.00	3.0
(5,5)	.00	4	.00	5	.00	.00	.00	9.0
(6,0)	.00	1	.00	.00	.00	.00	.00	1.0
(6,2)	.00	5	8	3	1	.00	.00	17.0

表 1　LVQ マップの例

4. 実　　験

我々は実験から、伝播学習が分類器の学習能力に及ぼす影響を示し、TaxSOM のマップに現れるクラスタが分類器の位置関係と一致することを検証する。

4.1 Reuter 新聞記事データ

本稿では Reuter 新聞記事データで実験を行う。文書総数は 20000 件, 総単語数約 240 万個の XML 形式のデータを使用する。総クラス数は 126 件で、うち全体の約 10%である 2000 件以上の記事が属する主だった 7 つのクラスについて実験を行う。この 7 つのクラスは図 3 のような分類階層構造を持っている。全ての記事データは少なくともひとつのトピックにあらかじめ分類されている。クラスカテゴリとその配分を表 2 に示す。各クラスの分布の総和が 20000 件を越えるのは複数クラスに属する文書が多数存在するからである。

No	Code	Description	Distribution
1	C15	performance	3678
2	C151	accounts/earning	2195
3	CCAT	corporate/industrial	9386
4	ECAT	economic	3038
5	GCAT	government/social	6181
6	M14	commodity markets	2287
7	MCAT	markets	5087

表 2　データのクラス分布

4.2 データの数値化

我々は 7 つのクラスに対する文書内の各単語の重要度によって文書データを数値ベクトル化する。この文書ベクトル x を特徴ベクトルと呼び、式 5 で定義する。

w_i は各文書中の i 番目の単語であり，n は各文書中の総単語数を示す。t_j は文書 x が属するクラス ($j = 1, 2, ..., 7$) である。$Freq(w_i, t_j)$ は，クラス t_j に属する文書中で単語 w_i が出現する頻度を示し，$Freq_{w_i}$ は全文書中での単語 w_i の出現頻度を示す。各単語の重要度を出現頻度を利用して算出しその総和をとることで，クラス t_j における文書 x の重要度を示している。

まず、与えられた文書に対して形態素解析 [10] を行い，文章を意味のある (効果的な) 単語単位に分割する。例えば, 我々は

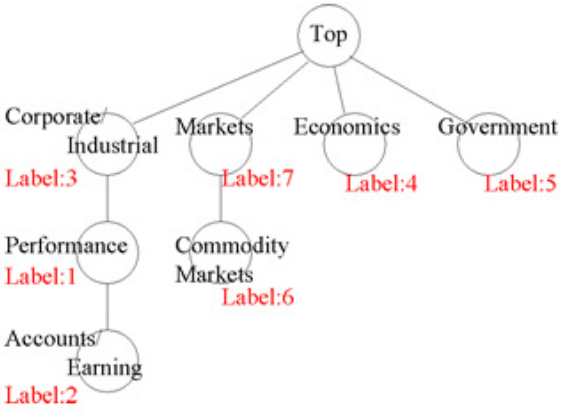

図 3　分類階層構造

「Blue chips end up as Fed keep interest rates steady」という文章を「Blue/chips/end/up/as/Fed/keep/interest/rates/steady」と分割することができる。さらに，これらのなかから特に分類に対して効果的に働く単語を何らかの基準によって抽出する。一般的に，文書データは基準によって抽出した単語の出現頻度による数値ベクトル表現 (文書ベクトル) に変換されるが，この手法では出現頻度が低くても分類に強く影響する単語を考慮できない問題がある。本稿ではある文書に現れる単語 w_i が分類に使用する全単語中でどの程度各カテゴリに影響を与えるかを考慮する。これにより 1 つの文書中の単語の出現頻度だけに依存しない重要度による数値化を行う。

$$x(w_i, t_j) = \sum_{i=1}^{n} \left(\frac{Freq(w_i, t_j)}{Freq_{w_i}} \times \ln \frac{Freq_{w_i}}{Freq(w_i, t_j)} \right) \qquad (5)$$

4.3 実験方法

実験は 10000 件づつ 2 つの集合に分けて行う。半分を分類ルール作成のために訓練データとして利用し、残りをルールの評価のために試験データとして利用する。記事の重複はなく、従来の TaxSOM と TaxSOM(k) で同様の実験を行い、再現率、適合率によるマルチクラスデータに対する分類器としての評価を行う。TaxSOM(k) の伝播数は $k = 1 \sim 10$ の間で行う。訓練データから得られた学習結果で試験データを分類し、訓練マップ上に配置された各点の試験データのクラスと、学習で得た各点のクラス分布のなかで支配的に働くクラスとを比較する。

4.4 評価方法

TaxSOM(k) の分類は各点上で支配的に働くクラスをその点のクラスとし、再現率 (式 6) と適合率 (式 7) を求める。一般的にこの 2 つの値はトレードオフの関係を持っているが、どちらか一方が極端に低い場合、分類器としての利用価値を見出すことはできない。本稿では両方がバランスよく成り立つポイントを測定するために再現率と適合率から F 値 (式 8) を計算し、判断基準とする。また、マップ上の各点ごとに正答率 (式 9) を求めて各点での分類能力を測る。各点に振り分けられた試験データのクラスと各点のクラス分布中で支配的に働くクラスが一致するならば、その分類を正答とみなす。複数クラスを持ったデータの場合は少なくともひとつが正解クラスに含まれていれば正答とする。仮にラベルが 45 の点にクラス 4 または 5 を含むデータが分類された場合、この分類は正答であると考える。また、分類階層構造を考慮していることから、上位クラスのラベルを持つ点に下位クラスのデータが分類されるときも正解とみなす。

$$\text{再現率} = \frac{(\text{マップ上に配置されたデータ総数})}{(\text{試験データ総数})} \qquad (6)$$

$$\text{適合率} = \frac{(\text{マップ上に正しく配置されたデータ総数})}{(\text{マップ上に配置されたデータ総数})} \qquad (7)$$

$$F - measure = \frac{(2 * \text{適合率} * \text{再現率})}{(\text{適合率} + \text{再現率})} \qquad (8)$$

$$\text{正答率} = \frac{(\text{マップ上の各点に正しく配置されたデータ総数})}{(\text{マップ上の各点に配置されたデータ総数})} \qquad (9)$$

5. 実験結果

5.1 結　　果

従来の TaxSOM による分類と、TaxSOM(k) による分類の結果を示す。伝播数 k は $k = 2, 3, 5, 10$ について実験結果を示す。それぞれのクラス分布と各点の正答率の結果を表 7~9 に示す。図 4~8 にそれぞれの出力マップを示す。

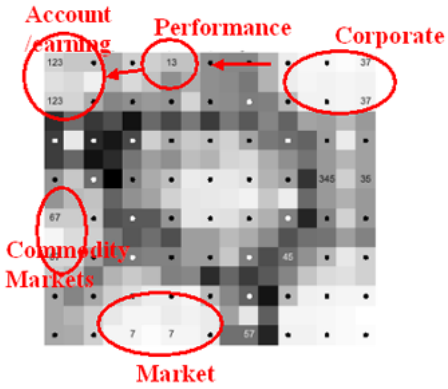

図 4　TaxSOM マップ

Point	ラベル	データ総数	正答数	正答率
(0,0)	123	784	563	71.81%
(0,1)	123	269	211	78.44%
(0,4)	67	469	297	63.33%
(0,5)	67	152	143	94.01%
(2,7)	7	745	403	54.09%
(3,7)	7	693	447	64.50%
(5,7)	57	28	22	78.57%
(6,5)	45	334	207	61.98%
(7,3)	345	812	485	59.73%
(8,0)	37	107	79	73.83%
(8,1)	37	51	45	88.24%
(8,3)	35	651	437	67.13%
	合計	5095	3339	65.53%

表 3　TaxSOM:各点のラベルと正答率

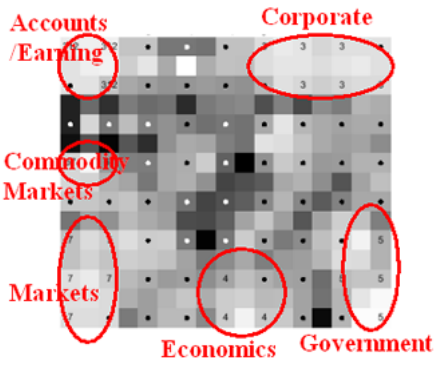

図 5　TaxSOM(2) マップ

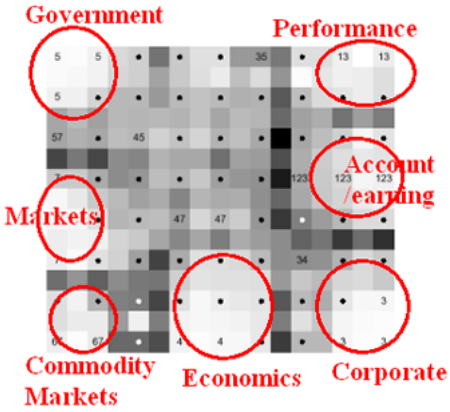

図 6　TaxSOM(3) マップ

Point	ラベル	データ総数	正答数	正答率
(0,0)	312	492	257	52.24%
(0,3)	67	542	246	45.39%
(0,5)	7	951	431	45.32%
(0,6)	7	463	258	55.72%
(0,7)	7	13	12	92.31%
(1,0)	312	126	58	46.03%
(1,1)	312	211	103	48.82%
(1,3)	67	86	64	74.42%
(1,5)	7	23	16	69.57%
(1,6)	7	90	78	86.67%
(4,6)	4	259	139	53.67%
(4,7)	4	331	164	49.55%
(5,0)	3	69	66	95.65%
(5,7)	4	45	32	71.11%
(6,0)	3	21	17	80.95%
(6,1)	3	644	395	61.34%
(7,0)	3	125	103	82.40%
(7,1)	3	21	21	100%
(7,6)	5	33	29	87.88%
(8,1)	3	16	16	100%
(8,5)	5	72	59	81.94%
(8,6)	5	77	46	59.74%
(8,7)	5	10	10	100%
	合計	4720	2620	55.51%

表 4　TaxSOM(2):各点のラベルと正答率

Point	ラベル	データ総数	正答数	正答率
(0,0)	5	613	472	77%
(0,1)	5	232	121	52.16%
(0,2)	57	260	148	56.92%
(0,3)	7	322	265	82.30%
(0,4)	7	561	373	66.49%
(0,5)	7	79	65	82.28%
(0,7)	67	52	51	98.08%
(1,0)	5	366	263	71.86%
(1,7)	67	649	568	87.52%
(2,2)	45	328	177	53.96%
(3,4)	47	111	59	53.15%
(4,4)	47	348	174	50%
(5,0)	35	293	155	52.9%
(6,3)	1237	15	15	100%
(6,5)	34	385	181	47.01%
(7,0)	13	233	215	92.28%
(7,3)	123	193	191	98.96%
(7,7)	3	797	662	83.06%
(8,0)	13	399	233	58.4%
(8,3)	123	13	13	100%
(8,6)	3	31	26	83.87%
(8,7)	3	75	74	98.67%
	合計	6355	4501	70.83%

表 5　TaxSOM(3):各点のラベルと正答率

5.2　考　　察

TaxSOM による分類の結果 (表 3, 表 9, 図 4) を基本として、伝播を進めたときの結果の変化をみる。TaxSOM(2) の結果 (表 10,4, 図 5) を見ると、伝播なしのときよりも分類点として機能する点が増加している。また, 伝播前は現れなかった Economics と Government の集合が現れている。しかし、分類不可能と判定され、どの点にも配置されなかったデータ数が増加している。再現率・適合率 (表 8) をみてもどちらも低下していることがわかる。マップを比較すると、2 回伝播を行った後の方がクラスタが見えなくしまうが、同じクラスが近い位置に集まっているのがわかり、トポロジを保っていることが図 5 から確認できる。

分類不可能と判定されたデータ群を確認すると、階層の最上位のクラスを複数持っているデータがほとんどであった。このことから伝播数 2 のマップでは複数のクラスに属するデータを判定できないことがわかった。

伝播数 3 回の結果 (表 5, 表 11, 図 6) は全体のなかで最も適合度 (表 8) が高い。再現率も伝播なし、2 回伝播時よりも高くなって約 63%となる。2 回伝播時との大きな差は、複数の最上位クラスに属するデータが分類できるようになっている点である。マップを見ると、TaxSOM よりもクラスタが細分化されてマップ上に現れている。また,2 回伝播時と比較してクラスタがはっきりあらわれ、マップ右側と左下側に階層構造 (図 3 参照) の影響をみることができる。分類不可能と判定されたデータは

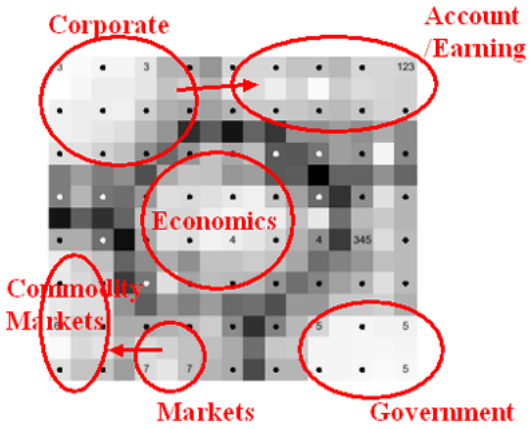

図 7　TaxSOM(5) マップ

Point	ラベル	データ総数	正答数	正答率
(0,0)	3	1942	1254	64.57%
(0,6)	67	481	418	86.9%
(2,0)	3	192	172	89.58%
(2,7)	7	642	498	77.57%
(3,7)	7	1425	732	51.37%
(4,4)	4	421	367	87.17%
(6,4)	4	201	83	41.29%
(6,6)	5	174	84	48.28%
(7,4)	345	47	44	93.62%
(8,0)	123	436	396	90.83%
(8,6)	5	98	81	82.65%
(8,7)	5	1388	768	55.33%
	合計	7447	4897	65.75%

表 6　TaxSOM(5):各点のラベルと正答率

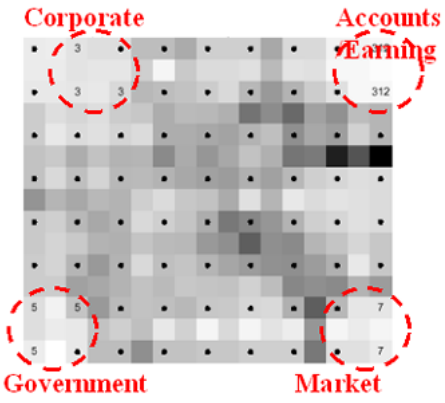

図 8　TaxSOM(10) マップ

様々だが明確な特徴はみつからない。

5 回伝播学習 (表 6, 表 12, 図 7) では、分類点として働く数は減少しているが、再現率が最も高くなっている。各点の分類結果とマップ (図 7) をみると、3 次伝播次に比べると, 上位クラス同士の境界線 (例えば Economics と Government の間) が明確に

Point	ラベル	データ総数	正答数	正答率
(0,6)	3	328	275	83.84%
(0,7)	67	2524	1506	59.67%
(1,0)	3	1589	824	51.86%
(1,1)	7	239	201	84.1%
(1,6)	7	187	158	84.49%
(2,1)	4	156	114	73.08%
(8,0)	4	948	791	83.44%
(8,1)	5	281	192	68.33%
(8,6)	345	344	267	77.62%
(8,7)	123	643	423	65.76%
	合計	7239	4751	65.63%

表 7　TaxSOM(10):各点のラベルと正答率

Classifier	データ数	正答数	再現率	適合率	F 値
TaxSOM	5095	3339	50.95%	65.53%	57.33%
TaxSOM(2)	4720	2620	47.2%	55.51%	51.02%
TaxSOM(3)	6355	4501	63.55%	70.83%	66.99%
TaxSOM(5)	7447	4897	74.47%	65.76%	69.84%
TaxSOM(10)	7239	4751	72.39%	65.63%	68.84%
SVM-LinearE(50)			48.89%	48.89%	48.89%
SVM-LinearU(30)			48.89%	53.66%	53.66%
SVM-Gauss-E(30)			37.78%	53.13%	44.16%
SVM-Gauss-(20)			37.78%	53.13%	44.16%

表 8　F 値による分類能力の評価

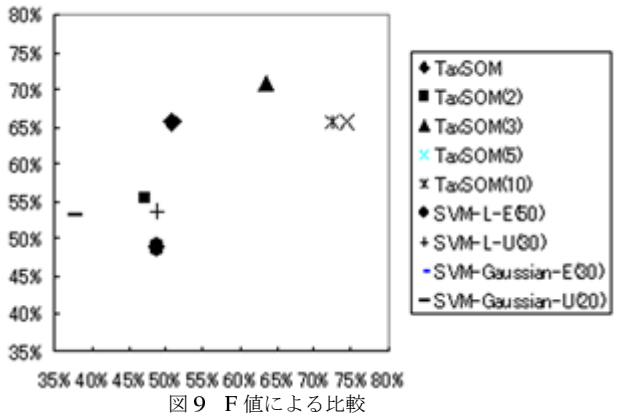

図 9　F 値による比較

なっているのに対して, 上下関係を持つクラス間 (例えば Market と Commodity Market) の境界線は薄くなっていることから、下位クラスと上位クラスの集約が始まっていると考えることができる。また、各クラス同士の関係をマップ上のクラスタの位置から推測することができるのも興味深い点である。ここで F 値 (表 8) が最も高くなっていることから、分類器としては 5 回伝播の状態が最もよい結果を示したと判断できる。

10 回伝播 (表 7, 表 13, 図 8) では 5 回伝播の結果よりもさらに集約が進んで、上位クラスへのクラス分布が明確になっている。下位クラスに属するデータも上位クラスの分類点に配置されていることから、伝播学習によって学習が進められた結果、下位データの特徴を含むコードブックの生成に成功していると考えられる。Economics を明確に分類する点が消え, このクラスに属するデータが他のクラスの分類点に分類されるように

なったことが興味深い点である。このことからEconomicsクラスは他のクラスの特徴をすべて含んだクラスであったと考えられる。しかし、再現率、適合率(表8)は5回伝播よりも減少する結果となっている。また図8からクラスタが完全に読み取れなくなっていることから,過度の伝播はマップに悪影響を及ぼすと考えられる。

本研究で得た結果を、Luis Gravano [5] らの研究のWebページのSVMを分類器とした場合に得られている結果と比較する(表8)。様々な研究でSVMの分類能力の高さが検証されているが，Gravanoらの研究からマルチクラスの分類に関して,SVMの分類能力は効果的に働かない場合があることを確認できる。図9に表8で示した再現率，適合率を用いた比較を示す．縦軸で適合率，横軸で再現率をあらわし，右上角に近づくものほど，その手法が分類器として有効であると考えられる．これらの比較から，本稿の手法が良好な結果を示していることがわかる。

6. 結　　び

本研究から、我々が提案するTaxSOM(k)モデルがマルチクラスデータの分類器として有効に機能することを示した。このモデルによって、従来のSOMで失っていた文書データのトポロジ問題を解決することが可能となった。また、クラスタ化能力を保ちながら,マップ上の各点を分類器として機能させることで,クラスタ化とクラス分類の同時出力を実現できた。結果から過度の伝播学習がマップに悪影響を及ぼす可能性がみられたが、現時点ではどの程度の伝播数がトポロジ、分類能力に最も良い影響を与えるかを事前に知る方法がない。伝播数の目処をつける何らかの手法を発見することが今後の課題となる。また，本研究で設計したTaxSOM(k)は我々の研究室のホームページによりフリーウェアとして公開予定である．(http://www.dbl.k.hosei.ac.jp)

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14580392)の支援による．

文　　献

[1] G.Adami,P.Avesani,D.Sona. "Bootstrapping for Hierarchical Document Classification" *Conference on Information and Knowledge Management* (CIKM),2003

[2] G.Adami,P.Avesani,D.Sona. "Clustering documents in a web directory" *Workshop On Web Information And Data Management*,2003

[3] T.Joachims.: "Text Categorization with Support Vector Machines", *Proc.European Conf. on Machine Learning* (ECML),1998

[4] コホネン.T, : 自己組織化マップ , シュプリンガー・フェアラーク東京,1996

[5] L.Gravano,V.Hatzivassiloglou,R.Lichtenstein. "Categorizing Web Queries According to Geographical Locality" *Conference on Information and Knowledge Management* (CIKM),2003

[6] T.Miura, T.Yanagida: "k-propagated Self-organizing Maps" *Artificial and Computional Intelligence* (ACI),2002

[7] 柳田 卓郎, 三浦 孝夫: "k 次伝播 SOM によるデータ分類", *Data Base Work Shop*(DBWS), 2002

[8] T.Miura, T.Yanagida,Isamu SHIOYA: "Classifying News Corpus by Self-Organizing Maps" *IEEE Pacific Rim Conference on Communications, Computers and Signal processing* (PACRIM),2003

[9] Y.Yang and J.Pedersen. "A comparative sutudy on feature selection in text categorization" *International Conference on Machine Learning* (ICML),1997

[10] 長尾 真: 自然言語処理, 岩波書店, 1996

[11] 永田, 平田.: "テキスト分類–学習理論の「見本市」–", 情報処理,vol.42(1),pp:32–37(2001)

[12] S.Wermter, Chihli Hung: "Selforganizing classification on the Reuter news corpus", *The 19th International Conference on Computational Linguistics* (CO LING),2002

付録：各手法のクラス分布表

Point	1	2	3	4	5	6	7
(0,0)	.0485	.4477	.3456	.0025	.1326	.0001	.0229
(0,1)	.0892	.5018	.1895	.0334	.0520	.0334	.1003
(0,4)	.0043	.0001	.0511	.0001	.0980	.4029	.4434
(0,5)	.0001	.0001	.0511	.0001	.0001	.6315	.3092
(2,7)	.0116	.0083	.0592	.0001	.0001	.1564	.7703
(3,7)	.0346	.0158	.1471	.0505	.1067	.0735	.5714
(5,7)	.0001	.0001	.0001	.0001	.3181	.0001	.6818
(6,5)	.0001	.0001	.0001	.4326	.5673	.0001	.0001
(7,3)	.0001	.0001	.1711	.2894	.3842	.0529	.1022
(8,0)	.0001	.0001	.2616	.1495	.0001	.0001	.5887
(8,1)	.0980	.0588	.2745	.0392	.0784	.0980	.3529
(8,3)	.0246	.0583	.3502	.0001	.4854	.0491	.0322

表 9　TaxSOM:クラス分布

Point	1	2	3	4	5	6	7
(0,0)	.0873	.6605	.1890	.0020	.0243	.0162	.0203
(0,3)	.0001	.0001	.0129	.0258	.0018	.4631	.496
(0,5)	.0094	.0126	.0147	.0325	.0115	.0389	.8801
(0,6)	.0001	.0001	.0215	.0410	.0086	.1166	.8120
(0,7)	.0001	.0001	1.000	.0001	.0001	.0769	.9230
(1,0)	.2063	.5158	.2063	.0714	.0001	.0001	.0001
(1,1)	.1753	.5876	.2180	.0001	.0142	.0001	.0047
(1,3)	.0116	.0813	.0465	.0930	.0232	.5930	.1511
(1,5)	.0001	.0001	.0001	.0001	.0001	.0001	1.000
(1,6)	.0555	.0777	.0001	.0001	.0001	.0001	.8666
(4,6)	.0501	.0308	.0193	.7606	.0270	.0308	.0810
(4,7)	.0030	.0030	.0001	.9607	.0302	.0001	.0030
(5,0)	.0001	.0001	.9565	.0289	.0144	.0001	.0001
(5,7)	.0001	.0001	.0444	.8888	.0001	.0444	.0222
(6,0)	.1904	.0001	.8095	.0001	.0001	.0001	.0001
(6,1)	.0481	.0010	.9239	.0155	.0001	.0001	.0015
(7,0)	.0720	.0160	.8240	.0880	.0001	.0001	.0001
(7,1)	.0001	.0001	1.000	.0001	.0001	.0001	.0001
(7,6)	.0001	.0001	.0001	.1212	.8787	.0001	.0001
(8,1)	.0001	.0001	1.000	.0001	.0001	.0001	.0001
(8,5)	.0001	.0001	.0001	.1250	.8194	.0555	.0001
(8,6)	.0001	.0001	.0001	.1688	.7922	.0389	.0001
(8,7)	.0001	.0001	.0001	.0001	1.000	.0001	.0001

表 10　TaxSOM(2):クラス分布

Point	1	2	3	4	5	6	7
(0,0)	.0001	.0212	.1076	.0032	.7699	.0505	.0114
(0,1)	.0689	.0431	.0732	.1034	.5215	.0818	.1077
(0,2)	.1741	.0393	.0056	.0561	.3426	.0001	.3820
(0,3)	.0279	.0062	.0900	.0341	.0186	.0001	.8229
(0,4)	.0124	.0001	.2620	.0001	.0001	.0071	.7183
(0,5)	.0126	.0001	.1645	.0001	.0001	.0001	.8227
(0,7)	.0001	.0001	.0011	.0001	.0001	.4222	.5666
(1,0)	.0382	.0601	.0846	.0601	.7185	.0109	.0273
(1,7)	.0001	.0001	.0019	.0392	.0001	.4007	.5579
(2,2)	.0485	.0001	.1119	.3768	.4626	.0001	.0001
(3,4)	.0001	.0001	.0001	.2558	.3430	.0232	.3779
(4,4)	.0001	.0129	.0001	.3126	.1576	.0077	.5090
(5,0)	.0472	.0292	.5202	.0001	.4031	.0001	.0001
(6,3)	.2000	.2000	.3333	.0001	.0001	.0001	.2666
(6,5)	.0285	.0389	.3532	.3038	.1038	.1298	.0415
(7,0)	.6008	.0001	.3218	.0257	.0300	.0171	.0042
(7,3)	.0984	.6165	.2746	.0001	.0103	.0001	.0001
(7,7)	.0001	.0001	.8306	.0602	.0058	.0213	.0288
(8,0)	.2972	.1133	.2846	.1360	.0237	.0579	.0780
(8,3)	.1538	.3076	.5384	.0001	.0001	.0001	.0001
(8,6)	.0001	.0001	.8387	.0001	.0967	.0322	.0322
(8,7)	.0001	.0001	.9866	.0001	.0133	.0001	.0001

表 11　TaxSOM(3):クラス分布

Point	1	2	3	4	5	6	7
(0,0)	.0782	.0777	.4987	.1096	.0849	.0865	.0731
(0,6)	.0478	.0228	.0311	.0124	.0166	.6403	.2286
(2,0)	.0001	.0520	.8437	.0416	.0625	.0001	.0001
(2,7)	.0560	.0529	.0560	.0001	.0591	.0529	.7227
(3,7)	.1066	.1480	.2231	.0001	.0001	.1719	.3501
(4,4)	.0001	.0001	.0047	.8717	.0807	.0001	.0427
(6,4)	.0497	.0746	.1343	.4129	.0001	.0001	.3283
(6,6)	.0977	.0287	.3160	.0001	.4827	.0057	.0689
(7,4)	.0212	.0425	.4680	.1489	.3191	.0001	.0001
(8,0)	.1146	.3600	.4334	.0001	.0412	.0001	.0504
(8,6)	.0306	.0102	.0918	.0408	.8265	.0001	.0001
(8,7)	.0001	.0007	.0835	.0670	.7478	.0001	.1008

表 12　TaxSOM(5):クラス分布

Point	1	2	3	4	5	6	7
(0,6)	.0001	.0001	.0094	.0670	.8384	.0001	.0001
(0,7)	.0154	.0162	.0412	.0265	.7995	.0095	.0915
(1,0)	.0648	.1428	.6910	.0075	.0604	.0132	.0201
(1,1)	.0460	.0585	.8410	.0001	.0543	.0001	.0001
(1,6)	.0001	.0001	.0106	.0001	.8449	.0001	.1443
(2,1)	.0192	.0705	.7756	.0001	.0001	.0001	.1346
(8,0)	.1455	.4345	.2542	.0001	.1371	.0052	.0232
(8,1)	.0441	.7977	.0073	.0001	.1139	.0183	.0183
(8,6)	.0001	.0001	.0203	.1337	.0116	.0581	.7761
(8,7)	.0202	.0575	.0793	.0015	.0279	.1010	.7122

表 13　K=10 のクラス分布