DEWS2003 7-C-01

SOMによるテキスト分類

柳田　卓郎† 　三浦　孝夫† 　塩谷　勇††

† 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
†† 産能大学 経営情報学部　〒259–1197 神奈川県伊勢原市上粕屋 1573
E-mail: †{i02r3244,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし　本稿は文書データの自動分類手法としての *Self-Organizing Map*(SOM) [1] について述べる. 我々は,SOM の近傍学習を拡張し SOM マップから得られる情報をより明確にすることのできる *K* 次伝播 SOM(*K-SOM*) [7] を提案している.K-SOM を用いた分類では SOM マップ上の各点がクラス分布を保持することで, ある点に振り分けられたデータに対する詳細なクラス情報を得る事が可能になる. 本研究では,Reuter 新聞記事データを使った SOM と K-SOM によるテキスト分類を行い, その分類能力の有効性を評価する.

キーワード　テキスト分類,SOM,K 次伝播 SOM

Text classification by using Self-Organizing Maps

Taqlow YANAGIDA†, Takao MIURA†, and Isamu SHIOYA††

† Dept.of Elect.& Elect. Engr. HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
†† Department of Management and Information Science, SANNO University　1573, Kamikasuya, Isehara city, Kanagawa 259–1197 Japan
E-mail: †{i02r3244,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract　This paper purpose *Self-Organizing Map*(SOM) for Text classification. By using *K-propagated SOM*(K-SOM), we can find class membership distribution of each point on SOM map. In this work, we try to classify of Reuter news corpus by using SOM and K-SOM. And we estimate the performance of SOM and K-SOM as classifiers.

Key words　Text classification,SOM,k-propagated SOM

1. はじめに

ここ数年でインターネットが爆発的に普及したことに伴って, これまで紙とファイルに収められていた過去の様々なデータが電子化される動きが活発になっているが, それらは非常に大きなデータ量のものが多く, 利用者が有意義にデータを利用することが困難になっている. 文書データの分類はその文書を読んで内容を把握して行われるのが理想的だが, 数年分の新聞記事を人間が読んでひとつ仕分けるのは気の遠くなるような作業であるため, 様々な手法で自動分類の試みがなされている.

文書データの自動分類は, 各単語の出現頻度を用いた高次元ベクトルを用いて行わ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. また, 高次元データに対する学習は出力結果が難解であることが多く, 解釈が難しい [2].

従来の次元縮小の手法として, 主成分分析や特異値分解による次元縮小をあげることが出来るが, 情報の要約による精度の低下が問題となる. この問題を解決するために, あらかじめ文書に与えられたクラスに対する単語の重要度を用いて数値化を行うベクトル表現 [6] が提案されている. これによって精度の低下を抑えた大幅な次元縮小が可能となる. 出力結果の解釈についてはコホネンの SOM マップを用いることで 2 次元マップに可視化することが可能であり, 他の手法に比べて SOM の出力は非常に有効であることが知られている. しかし, マップに対する説明不足から曖昧な部分も多く, 偶然の結果である可能性を払拭することが出来ない.

我々は既に SOM の近傍学習を拡張し, 学習によって得たクラス分布を元に SOM マップに対する客観的評価 [7] を行う手法を示している.

本研究では,SOM による文書データの分類について論じる.K-SOM の伝播学習で得られるクラス分布に, 適合度検定で信頼度を算出し,SOM マップの分類能力の有効性を検証する.

2 章では文書データの分類とその準備について述べる.3 章では分類手法と結果の評価法について論じ,4 章で文書データの数値化手法について論じる. 5 章では実験結果と考察を示し, 6 章で結びとする.

2. 文書データの分類

テキスト分類はテキスト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 [4]. 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. 自動分類には文書に対する何らかの分類ルールが必要で, テキスト分類では自然言語処理のための形態素解析 [3] と教師あり学習の併用によって行われる. 教師あり学習のための訓練データには記事データとその記事の正解クラスが与えられ, 分類器はデータの特徴とそのクラスの相関を学習する. 形態素解析は文章を意味のある単語単位に分割する手法である. 例えば「Blue chips end up as Fed keep interest rates steady」という文章は「Blue/chips/end/up/as/Fed/keep/interest/rates/steady」のように分割することができる. この単語の中から特に分類に重要な役割を果たす単語を何らかの基準によって抽出する. これを属性選択と呼ぶ. 属性選択によって抽出された単語が文書中で何回現れるかを数え, その値を属性値として文書をベクトルで表現する方法が一般的である. このベクトルとカテゴリとの対応関係について, 訓練事例を使って教師あり学習を行う. 文書データの自動分類は, 各単語の出現頻度を用いた高次元ベクトルを用いて行わ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. 今のところ Support Vector Machine(SVM) [5] 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ことが検証されているが, ルールによる分類ではなく, ベクトルとクラスの対応関係は複雑な数式によってあらわされるため結果の解釈が難しい.

2.1 SOM による分類

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 次元マップに可視化することができる。マップはトポロジ情報を持っていてマップ上の近さがデータの近さをあらわしクラスタを視覚的に把握できる。マップ上の点には支配的に働いたデータのクラスラベルが与えられる。SOM を分類器とみなす時、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 件割り当てられた時, 点にはラベル 1 が与えられる. また、マップの判断が実験者の主観に依存するため結果に対する信頼性の問題がある. 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. 各点に振り分けられた試験データのクラスと SOM マップ上のラベルが一致するならば、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ラベル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 1 の点にクラス 1,3 を持つ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。SOM マップの出力例を図 1 に示す.

SOM の学習を利用した分類器に LVQ がある.LVQ は教師あり競合学習によって訓練を行い, 訓練結果を使って未知のデータに対するクラス分類を行う。データ同士の距離が近く、与えられたクラスが同じであるものを 2 次元平面上の近い位置に配置していく。SOM との相違点は配置される時トポロジが考慮されない点で、LVQ から得られる情報は可視化された大まかなクラスの集合である。LVQ も SOM と同様に結果の判断が実験者の主観に依存する問題がある。

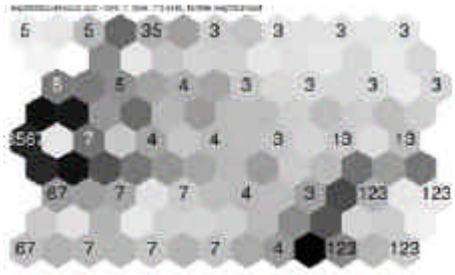

図 1 SOM マップの例

2.2 K-SOM による分類

K-SOM は SOM の近傍学習を拡張した教師あり競合学習を行う。SOM の学習によって得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。各点は伝播されたクラス情報を分布として保持する。この拡張によって、K-SOM は LVQ と SOM の持つ問題点を解決し, マップはトポロジ情報を保持しながら複数のクラスをまたぐ特徴を持ったデータ分類ができるようになる. マップ上の各点のクラス分布を表す出力を LVQ マップと呼ぶ.LVQ マップの例を表 1 に示す. 結果の判断はクラス分布に対する適合度検定で行う. 適合度検定の詳細は後で述べる。

K-SOM の分類は各点で支配的に働くクラスをその点のクラスとして正答率を求め, 更に正解分布と K-SOM 学習で得たクラス分布の間で適合度検定を行い分布の信頼度を求める. マップの各点に対してクラス分布の信頼度を与えることで SOM の問題であるマップの曖昧さを消すことがこの手法の狙いである.

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

2.3 評 価 法

実験で得た結果に含まれる意味の解釈を行う際, 主観に頼っているだけでは他の実験結果との比較や有効性の判定ができない。実験者に依存せず定量的判断を下し, 同様の実験同士の結果の比較ができるようにするために, 全ての実験には何らかの基準に基づいた評価方法が必要である。SOM を使った実験の大半は誤分類率による評価を行っている。誤分類率は各点につ

いて分類の正否を問う手法である。扱うデータがあるクラスに属しているか否かという 2 値的な判断を下す場合には有効な手段である。しかし K-SOM で得られる分類結果はクラスの分布情報である。各点は決定的なひとつのクラスは持たないため振り分けられたデータに対する分類結果の正否を問うことは難しい。K-SOM の結果に対する評価はクラス分布そのものに対する評価であることが望ましい。適合度検定は実験で得た測定値の分布と正解から得られる期待値の分布の一致をとる。有意水準によってその分布の有効性を数値で評価することができる。この評価方法によって我々は SOM マップに対して実験者の主観から離れ, 曖昧さを失くした評価を下すことが可能となる。

適合度検定は 2 つの分布間で比較を行い分布が一致するかどうかを判定する. 式 1 で X^2 値を計算し, $X^2 < \chi^2$ ならば分布は一致すると考え, 検定に合格する. $X^2 > \chi^2$ ならば分布は一致しないと考え, 検定は不合格となる.

$$X^2 = \sum_{j=1}^{m} \frac{y_i - N_{pj}}{N_{pj}} \tag{1}$$

検定に値する点を *Hit-point* と呼び, *Hit-point* のうちで検定に合格する点を *Reliable-point* と呼ぶ. 合格する有意水準の値を信頼度と考え, これによりマップの有効性を客観的に評価することができる. ただし検定は点に振り分けられたデータが 10 件以上 (例では 5 件以上) ある時だけ行う. これは母数が少なすぎる場合, 検定の信頼性が低いことが知られているためである. 検定結果の例を表 2 に示す.

Point	X^2var	test
(2,1)	3.26	100%
(2,3)	16.27	95%
(5,5)	22.04	95%
(6,2)	1686.9	denied

表 2　適合度検定の例.95%：69.13, 100%:12.21

3. 文書データの数値化

3.1 数値化手法

文書データの数値化を考える. 属性選択によって抽出された単語が文書に含まれるか否かを 0 または 1 で表現することで文書を 2 値ベクトルであらわすのが最も単純な手法である. この手法は文書を数値化することはできるが, 文書中で単語が分類にどの程度の影響力を持つかを表現できない問題がある. ある文書中の単語の出現頻度を使ったベクトル表現は最も一般的であり,2 値ベクトルによる表現の問題を解決しているように見える. しかし出現頻度が低くても分類に強い影響を及ぼす単語が存在することを考慮すると, この手法でも不十分である. そこで我々はカテゴリ内での単語の分類への影響力を考慮した, 重要度によるベクトル表現を使う. これにより 2 値ベクトル表現の問題と, 出現頻度を使った表現の問題を解決することができる.

3.2 Reuter 新聞記事データ

本研究では Reuter 新聞記事データを使用して実験を行う. 実験に使ったのは XML 形式の総記事数 21,217 件, 総単語数 約 240 万個のデータで, 全てのデータは少なくともひとつのトピックが与えられている. 総トピック数は 126 件あるが, 実験では全体の約 10%である 2000 件以上の記事が割り当てられるような主だったトピック 7 つに対する分類を行う. 実験に使用するトピックとその配分を表 3 に示す.

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

表 3　クラスとデータの内わけ

3.3 文書データのベクトル表現

我々は主だった 7 つの既存のトピックに対する各単語の重要度をベクトル要素として利用する. この文書ベクトル x を特徴ベクトル (Significance Vector [6]) と呼び, 式 2 で定義する. また, 文書の数値化の例を表 4 に示す.

$$x(w,t_j) = \sum_{i=1}^{n} \Big(\frac{Frequency for word w_i into pict_j}{\sum_{j=1}^{m} Frequency for word w_i into pict_j} \times \ln \frac{\sum_{j=1}^{m} Frequency for word w_i into pict_j}{\sum_{i=1}^{n} Frequency for word w_i into pict_j} \Big) \tag{2}$$

Word	1	2	3	4	5	6	7
Blue	.00	.33	.00	.00	.67	.33	.00
chips	.00	.00	.00	.00	1.0	1.0	.00
end	.14	.29	.00	.00	.71	.71	.00
up	.30	.10	.10	.00	.50	.40	.00
as	.00	.00	.00	.00	.00	.00	.00
Fed	.00	.00	.60	.00	.40	.40	.00
keep	.00	.00	.50	.25	.50	.50	.00
interest	.00	.00	.57	.00	.29	.43	.00
rate	.00	.14	.71	.00	.29	.43	.00
steady	.00	.00	.57	.00	.43	.43	.00
sigvec.	.44	.86	3.1	.25	4.8	4.6	.00

表 4　特徴ベクトルの例

4. 実　　験

実験は訓練データ 10,000 件と試験データ 11,217 件に分けて行う. 訓練データと試験データへの振り分けはランダムに行い, 記事の重複はない.SOM,K-SOM で同様の実験を行って 7×5 のマップを生成し, マップ上の Hit-point の正答数, 信頼度を示し, 評価を行う.K-SOM の伝播数は 1〜10 の間で行う.

4.1 SOM による分類結果

正答数 (図 2) はどの点も 300 程度で, ほぼ全ての点で 90%を超える正答率 (図 3) を示している. 適合度検定の結果 (表 5) から点 (2,3) だけが 20%の信頼度を持ちそれ以外の点は信頼できない分類と判断する.

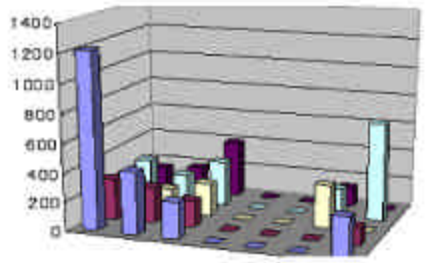

図 2　SOM の正答数

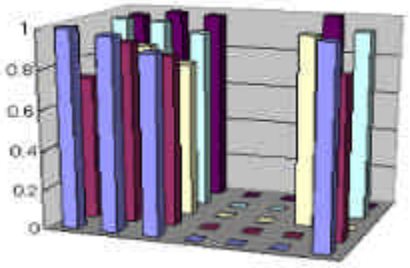

図 3　SOM の正答率

Point	X^2var	test
(0,0)	207.2	denied
(0,1)	400.2	denied
(0,2)	145.9	denied
(0,3)	318.2	denied
(0,4)	376.1	denied
(1,0)	299.2	denied
(1,1)	289.9	denied
(1,2)	220.5	denied
(1,3)	179.1	denied
(1,4)	376.1	denied
(2,0)	562.2	denied
(2,1)	464.2	denied
(2,2)	242.4	denied
(2,3)	126.9	20%
(5,0)	212.7	denied
(5,1)	388.5	denied
(5,2)	866.9	denied
(6,1)	589.0	denied
(6,3)	272.0	denied
(6,4)	682.9	denied

表 5　SOM の適合度検定 20%:139.1

4.2 K-SOM による分類結果

伝播数 K の増加とともに学習回数 (図 4) が減少し，正答数が増加していく. 正答率は実験中ほぼかわらず 90%程度，K=2 で信頼度 50%での検定合格数 (図 5) が最も高くなったあと，K の増加とともに合格数は減少していく. 例として伝播による各点の正答数と正答率の変化を K=2,5,10 を例として図 6 から図 11 に示す.

K=2 の時,Hit-point は 11 個現れる. 表 6 にクラス分布, 表 7 に検定結果と各点の正答数を示す.SOM の正答数と比較すると伝播によってデータが一部の点に集約していることがわかる. 特に (1,2) は 169→ 3255,(2,1) は 318→ 2452 と大幅に増加している. 検定結果から (1,2) は信頼度 95%の分類,(2,1) は信頼度 50%の分類と判断する. データに対する判断の一例を挙げると, 我々は (1,2) に属する 3255 個のデータは 95%の信頼度で, 主にクラス 3 と 5, つまり「産業」と「政治」(表 3 参照) について書かれた文書であると判断できる. 他の点でもクラス分布 (表 6) から, 点 (2,1) からクラス (3,5,7), 点 (3,1) からクラス (1,2,3) の相関をみることができる. Reliable-point の正答数 (表 7) の総和がマップ全体の正答数と考えると、K=2 時のマップ全体の正答率は信頼度 100%を満たす場合 $(654+10)/11217\times100 = 5.92\%$, 信頼度 95%を満たす場合 $(654+10+11+3255)/11217\times100 = 35.04\%$, 信頼度 50%を満たす場合 $(654+10+11+3255+814+1053+2452+503)/11217\times100 = 78.02\%$ となる.

Point	1	2	3	4	5	6	7
(0,0)	.00	.00	36	15	144	.00	11
(0,1)	.00	.00	.00	.00	2	.00	11
(0,2)	.00	.00	.00	9	.00	.00	.00
(1,0)	2	.00	36	15	79	5	19
(1,1)	36	11	79	2	.00	.00	.00
(1,2)	11	9	36	16	42	9	11
(2,0)	.00	.00	3	7	83	.00	.00
(2,1)	11	11	144	49	144	73	144
(3,1)	61	67	61	.00	.00	.00	.00
(6,3)	61	67	61	.00	.00	.00	.00

表 6　K=2 のクラス分布

Point	X^2var	reliability	correct
(0,0)	59.2	50%	814
(0,1)	27.9	95%	11
(0,2)	12.2	100%	10
(1,0)	11.7	100%	654
(1,1)	64.5	denied	367
(1,2)	13.2	95%	3255
(2,0)	53.5	50%	1043
(2,1)	45.8	50%	2452
(3,0)	180.3	denied	283
(3,1)	56.2	50%	503
(6,3)	376.0	denied	625

表 7 K=2 の適合度検定と各点の正答数 100%:12.2,95%:37.5,50%:59.3

K=3 の時, 表 8 にクラス分布, 表 9 に検定結果と各点の正答数を示す. K=2 の時と比較すると Hit-point の数が 2 減少するが, 各点の正答数は増加している. 正答数の増加が顕著だった点は (0,0) で 814→ 3077,(0,2) で 10→ 1259,(1,0) で 654→ 1880 となった. 各点の信頼度は表 9 に示す. (0,0),(2,0) では信頼度 50%→ 90%に向上しているが (1,0) は 100%→ 90%,(0,2) は 100%→ 40%,(1,2) は 95% → 40%に低下している. 表 6 と表 8 とを比較すると支配的に働くクラスへの比率が高くなっていることがわかる. k=2 の時と同様に Reliable-point の正答数 (表 9) からマップ全体の正答率を計算すると, 信頼度 90%の場合 51.13%, 信頼度 40%の場合 69.76%となる

Point	1	2	3	4	5	6	7
(0,0)	49	26	74	11	32	3	29
(0,1)	.00	.00	59	449	6	.00	6
(0,2)	4	27	133	75	133	8	85
(1,0)	1	.00	47	55	47	6	66
(1,2)	.00	.00	79	11	213	11	25
(2,0)	10	2	140	66	158	32	71
(3,0)	6	6	79	66	221	6	39
(3,1)	.00	.00	1	.00	542	.00	.00
(5,1)	.00	.00	.00	9	.00	.00	.00

表 8　K=3 のクラス分布

Point	X^2var	reliability	correct
(0,0)	59.2	90%	3077
(0,1)	27.9	denied	449
(0,2)	12.2	40%	1259
(1,0)	11.7	90%	1880
(1,2)	13.2	40%	813
(2,0)	53.5	90%	796
(3,0)	180.3	denied	387
(3,1)	56.2	denied	1055
(5,1)	376.0	–	11

表 9 K=3 の適合度検定と各点の正答数 100%:12.2,90%:41.2,40%:56.0

K=5 の時, 表 10 にクラス分布, 表 11 に検定結果と各点の正答数を示す. K=3 の場合と比較すると Hit-point の位置 (表 9,11) がほとんど変化しなくなったことがわかる. また K=3 の時よりもさらにはっきりと支配的に働くクラスがわかるようになっている（表 10) ただし伝播数の増加に伴って各点の正答数 (表 11 は増えている. マップ全体の正答率は信頼度 95%の場合 39.6%となる

K=7 の時, 表 12 にクラス分布, 表 13 に検定結果と各点の正答数を示す. K=5 の時とクラス分布 (表 12) を比較すると Hit-point の数がさらに減少している. 支配的に働くクラスの全体に対する比率が非常に大きくなっているのがわかる. 分布が極端に特徴化されすぎて Hit-point はひとつも検定 (表 13) に合格しなかった. よってこのマップは信頼できないものとして扱う.

Point	1	2	3	4	5	6	7
(0,0)	98	83	335	107	321	51	129
(0,1)	88	93	358	11	321	51	129
(0,2)	51	47	321	129	358	47	124
(1,0)	11	11	115	22	504	11	71
(1,1)	44	23	358	149	409	11	129
(1,2)	209	7	1598	10	17	34	41
(2,0)	47	4	207	56	24	1032	1099
(3,0)	.00	.00	3	20	2	.00	21
(3,1)	0	0	3	20	2	0	21
(5,4)	.00	.00	.00	9	.00	.00	.00

表 10　K=5 のクラス分布

Point	X^2var	reliability	correct
(0,0)	14.1	95%	1228
(0,1)	122.0	denied	1163
(0,2)	13.7	95%	124
(1,0)	36.3	95%	1482
(1,1)	42.0	95%	1389
(1,2)	2095	denied	1598
(2,0)	3377	denied	1999
(2,1)	741.1	denied	472
(3,0)	582.1	denied	331
(3,1)	36.2	95%	219
(5,1)	11.2	–	9

表 11 K=5 の適合度検定と各点の正答数 100%:12.2,95%:43.2,50%:59.3

Point	1	2	3	4	5	6	7
(0,0)	109	30	1407	109	134	15	35
(1,0)	215	49	527	89	32	920	1475
(1,2)	1	.00	2	78	2	209	962
(2,0)	4	.00	117	808	81	49	222
(2,1)	1710	1169	2371	9	17	.00	1
(3,0)	.00	.00	45	30	2033	.00	.00
(3,1)	5	1	500	440	848	12	38
(5,1)	.00	.00	.00	9	.00	.00	.00
(6.3)	.00	.00	.00	.00	.00	.00	5

表 12　K=7 のクラス分布

Point	X^2var	reliability	correct
(0,0)	1382	denied	1407
(1,0)	3369	denied	1477
(1,2)	1758	denied	962
(2,0)	1831	denied	808
(2,1)	7832	denied	2425
(3,0)	1591	denied	2033
(3,1)	502.6	denied	849
(5,1)	11.3	–	5
(6,3)	5.18	–	9

表 13　K=7 の適合度検定と正答数 100%:12.2,95%:33.1,20%:56.0

K=10 の時, 表 14 にクラス分布, 表 15 に検定結果と各点の正答数を示す.K=7 の場合とクラス分布 (表 14) を比較してみると分布がまるで別物になってしまっているのがわかる. これま

であまり変化がなかった Hit-point に (5,4) が加わり,95%の信頼度を持ったクラス (3,6,7) の集約された分類点であることがわかる.

マップ全体としての正答率は信頼度 95%の場合 0.017%となる.

Point	1	2	3	4	5	6	7
(0,0)	19	.00	147	758	42	978	2313
(1,0)	1697	1201	1756	8	9	.00	21
(2,0)	22	6	745	418	1042	93	177
(2,1)	116	3	577	111	1	107	169
(3,0)	191	39	1665	78	37	1	49
(3,1)	.00	.00	.00	9	,00	.00	.00
(4,1)	.00	.00	.00	9	.00	.00	.00
(5.4)	.00	.00	17	.00	.00	11	11

表 14　K=10 のクラス分布

Point	X^2var	reliability	
(0,0)	5607	denied	2314
(1,0)	3610	denied	1809
(2,0)	477.2	denied	1055
(2,1)	422.1	denied	578
(3,0)	2133	denied	1700
(3,1)	1470	denied	1979
(4,1)	13.3	–	9
(5,4)	25.9	95%	19

表 15　K=10 の適合度検定と正答数 100%:12.2,95%:43.2,50%:59.3

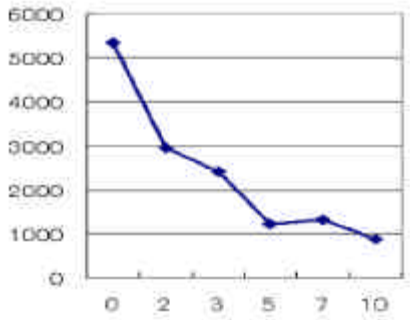

図 4　学習回数の推移

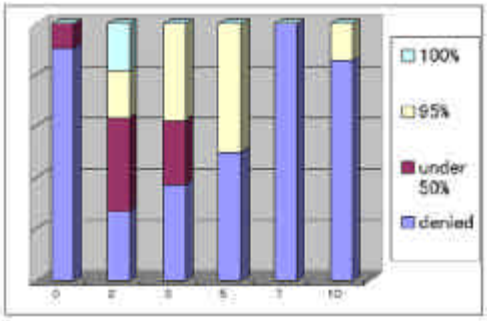

図 5　信頼度

4.3 考　　察

K=2∼7 の間はクラス分布に相関がみられたが k=10 になったときに大きく変わってしまった. これは過度の伝播学習による過学習が原因であると考えられる. 実験結果から最高で信頼性 50%の場合 78.02%の確率で分類予測が可能であることを示した. また、信頼性を重視した場合信頼度 90%で 51.13%の確率で分類予測が可能であることを示した.

5. 結　　び

本研究では K-SOM の学習によって SOM マップに信頼度の概念を導入し, 従来の主観的判断から数値による客観的判断への移行が可能であることを示した.

6. 付録：K-SOM の実験結果

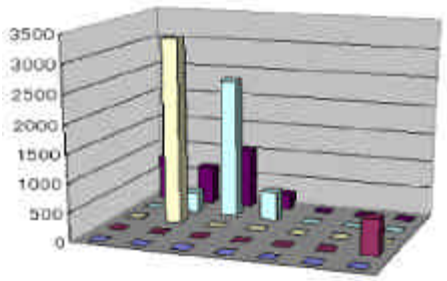

図 6　K=2 の正答数

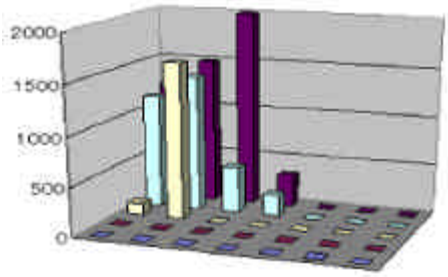

図 7　K=5 の正答数

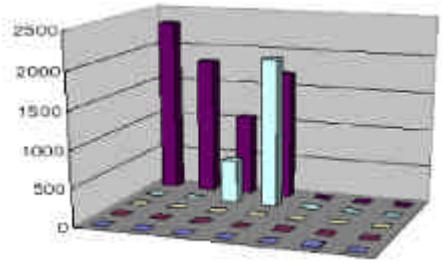

図 8　K=10 の正答数

文　　献

[1] Kohonen, T.: Self Organizing Maps, Springer-Verlag (1995)
[2] Y.Yang and J.Pedersen. ”A comparative sutudy on feature selection in text categorization” *International Conference*

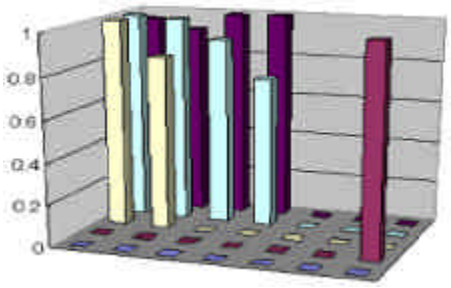

図 9　K=2 の正答率

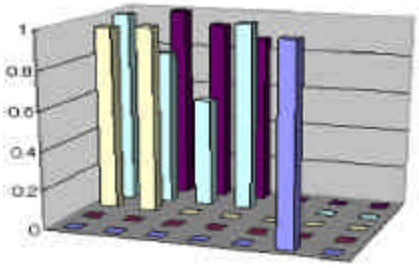

図 10　K=5 の正答率

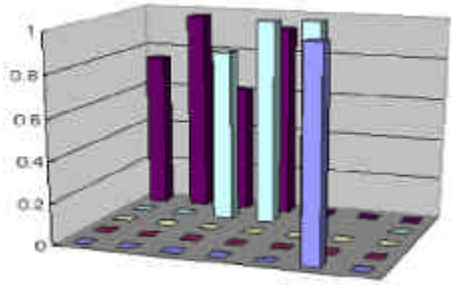

図 11　K=10 の正答率

on Machine Learning (ICML),1997

[3] **長尾 真**: **自然言語処理**, **岩波書店**, 1996

[4] **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)

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

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

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

[8] **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