SCONNによる文書の自動分類

3F－3

佐々木寛，羽生田博美，木下哲男
ＲＷＣＰ新機能沖研究室
{sasaki, haniuda, kino}@rwcp.or.jp

１．はじめに

近年，計算機ネットワークの発達に伴い，大量の文書情報が日々生成している．この様な情報の中には，FAQのように有用なものが多く存在し，これらを容易に検索・利用できるシステムが待ち望まれている．しかし，一般に検索効率は低く，目的の情報を取り出すことは容易でない．そこで検索効率を向上させる目的で，文書の自動分類をニューラルネットモデルの一つである特徴マップ[1]（Self-Organizing Map: 以下SOM)により行う手法が提案されている[2]．これによると，自動的に2次元平面上に文書地図を作成し，意味的に類似した文書同士に分類することが可能であるとの報告がなされている．

しかし，SOMではそのネットワーク構造が固定であるために，(a)構造決定の指針が立たないのではないか，(b)学習に多大の時間を必要とするのではないか等の問題が指摘される．

当研究室では，協調型システム研究[3]の一環として，ニューラルネットを利用し検索効率を向上させる種々の手法に関して研究を行っているが，本稿では，D.Choiらにより提案されたSCONN（Self-Creating and Organizing Neural Networks)モデル[4]に基づく文書の自動分類の手法を提案し，その性質と分類問題への適用性について議論する．

２．Self-Creating and Organizing Neural Networks

SCONNは教師なし学習を行うニューラルネットワークモデルの一つであり，入力層と必要に応じて徐々にノード数の増加する出力層からなる構造を持つ．これは，SOMの出力ノード数を動的に変化させるという改良を加えたものであり，出力ノードの学習性能を向上させている．

具体的には，まず入力層にN次元ベクトルxを提示すると，同次元のベクトルm_iを持つ出力ノードの内，ベクトルxに最も近いノードが勝者ノードとして選択される．次に，勝者ノードの活性度が活性度しきい値θ以下であるなら自身のベクトルm_iを修正し，それ以外は新たに勝者ノードの子ノードを生成し，それをm_iに近づける．これを繰り返すことで学習が進行し，最終的に出力ノードは入力ベクトル空間で活性度しきい値に応じた大きさのクラスタの代表ノードとなる．そこで，文書の分類問題においては，文書をベクトル化して入力すれば，最も近いベクトルを持つ出力ノードが分類先として選ばれ，自動分類が可能となる．

３．文書のベクトル化

文書をニューラルネットに入力するために，本稿では文書内に出現する基本単語の個数を文書のベクトル表現として用いる．

文書を単語の頻度を特徴としてベクトル化する場合，文書内に出現する全単語についてベクトル化することが望ましい．しかし，(1)学習にかなりの時間が必要となること，(2)少ない出現頻度を持つ単語が多いこと等から主要な単語を基本単語として予め人手で用意しておき，この単語だけについて頻度ベクトル化することとした．

この基本単語をもとに，文書の特徴となる単語を自動的に切り出し，これを次元とし，各要素が出現回数としたベクトル表現に変換する．

４．実験・考察

４-１．実験

本手法の有効性を検証するため，人手による分類

Classifying Articles Using SCONN
Hiroshi Sasaki, Hiromi Haniuda, Tetsuo Kinoshita
Real World Computing Partnership, Novel Functions Oki Laboratory
10-3, Shibaura 4-chome, Minato-ku, Tokyo 108, Japan

表1　各手法の動作条件

		SCONN	SOM
構造	IN	1006個	1006個
	OUT	上限100個	10×10個
学習回数		9000回	9000回
学習パラメータ		$\alpha(t) = 0.2\left(1 - \frac{t}{9000}\right)^2$	同左
活性度しきい値		$\theta(t) = 10\exp\left(1 - \frac{t}{2500}\right)$	——

およびSOMによる分類と比較する実験を表1の条件で行った．

分類対象としては，インターネットニュースにあるパソコンのハードウェアに関するニュースグループから500記事を使用した．本実験では，最頻出単語と出現頻度1の単語を除いた1006個を基本単語とした．人手でクラス分けをする際には，分類先のクラス数の上限を100までに抑えるよう条件付けた．

4-2．実験結果

各実験の結果，表2に示す分類結果が得られた．SCONNに基づく手法と，SOMを用いる手法共に多少の誤差はあるものの，意味的に類似した文書が各クラスに集められている．また，意味的に近いクラス同士が互いに近い位置に学習されていることも確認され，その結果，単語の出現パターンを用いてある程度の分類を行うことが可能であることが解る．

しかし，この分類が人間にとって，どの程度の意味的なまとまりを持つかは不明である．そこで上記の結果に基づき本手法と人による分類結果との比較を行った．まず，2つの分類状況の類似度を計る為に評価式（式1）を用意する．これは，人が分類したクラスCi内にあるNci個の記事を，本手法が他記事と誤

表2　分類状況

	クラス数	出力ノード数	学習時間
本手法	98	98	約2.5日
SOM	76	100	約10日
人間	98		

$$f(C_i) = \sum_{\text{for all } j} \frac{n_{Ci,\,Sj}^2}{N_{Ci} \cdot N_{Sj}} \quad \cdots\cdots \text{(式1)}$$

N_{Ci}: クラスCiに含まれる文書数(人間による分類)
N_{Sj}: クラスSjに含まれる文書数(アルゴリズムによる分類)
$n_{Ci,\,Sj}$: クラスCiとクラスSjに共通に含まれる文書数

表3　本手法と人との分類状況の比較(%)

97クラス（小分類）	22.98
36クラス（大分類）	26

表4　本手法とSOMとの分類状況の比較(%)

本手法による分類	22.98
SOMによる分類	15.31

らずにどの程度分類できたかを意味する．この評価式を用いて比較した結果を表3に示す．さらに，SOM法との比較も行うため，人による分類を基準にして，各々がどの程度，人による分類に類似しているかを比較した結果が表4である．

4-3．まとめ

以上の結果をまとめると

・SOMによる分類手法に比べ，学習時間が大幅に削減できる．

・SOMによる分類手法に比べ，分類精度に向上がみられた．これは本手法が，カテゴリ領域の複雑さに左右されない動的構造をとる為と思われる．

つまり，不定形・未知の大量データを自動分類する場合，学習時間と構造決定の不必要性および分類状況を考慮すると，パラメータ数が増えるとは言え，SCONNに基づく手法が有効と考えられる．

5．今後の方針

今回の実験では，入力データ表現として基本単語の共起パターンを用いた．今後は，基本単語の種類や数，そしてドメインの違いによる分類状況への影響について検討を進めていく予定である．

参考文献

[1] T.Kohonen, "The Self-organizing Map", pp.1464-pp.1480, Proc. of The IEEE, Vol.78, No.9, 1990.

[2] 有田英一，「AIによる問題解決技術の新展開-自己組織型情報ベース-」，1993年度電気関係学会東海支部連合大会．

[3] 羽生田博美, 木下哲男，「異種知識の協調利用に基づく情報検索システム」，情報処理学会第50回大会，3F-4, 1995.

[4] Doo-Il Choi and Sang-Hui Park, "Self-Creating and Organizing Neural Networks", IEEE Transactions on Neural Networks, pp.561-pp.575, Vol. 5, No.4, 1994.