ニューラルネットワークによる図書の自動分類

4T－10

畑田 稔
日立製作所 システム開発研究所

1. はじめに

インターネット時代に入り、小規模な企業図書館ではコンピュータ化による業務効率の向上、サービスの迅速化がより一層求められるようになった。

図書の管理では、日々受け入れる数多くの図書を整合のとれた形で分類することが必須である。分類方法としては、司書など専門家による国際十進分類法（UDC）、日本十進分類法（NDC）などが知られている。本研究では、自動分類の可能性を追求するために、書名からその書籍のNDCを決定する問題に階層型ニューラルネットワーク[1]を適用した。

2. 評価実験

2.1. 実験データ

電気、電子、通信、情報工学関連の5,500冊の和書をサンプルデータ（5,000冊を学習サンプル、500冊を評価サンプル）とした。分類法は日販のNDCを元に、発行年度による差異を無くした。

電気、電子、通信、情報工学はNDCでは、まず、表1に示すように10分野に分類される。

表1. 電気、電子関連のNDC（1次分類）

No	NDC	分野	件数
1	540	電気、電子、通信、情報全般	295
2	541	電気回路、計測、材料	192
3	542	電気機器	153
4	543	発電	255
5	544	送電、変電、配電	207
6	545	電灯、照明、電熱	93
7	546	電気鉄道	60
8	547	通信工学、電気通信	1,608
9	548	情報工学	1,199
10	549	電子工学	1,460
		合計	5,500

それぞれは更に最大11分類される。通信工学の分類方法を表2に示す。分野によっては、更に枝番号に分類されていくが、今回の評価実験では2次分類までを対象とした。

表2. 通信工学のNDC（2次分類）

No	NDC	分野	件数
1	547	通信工学、電気通信全般	147
2	547	通信機器・材料工業	63
3	547.1	通信回路、測定	31
4	547.2	通信方式、通信線路、通信網	197
5	547.3	通信機器・材料・部品、音響	181
6	547.4	有線通信	465
7	547.5	無線通信、電波、高周波	99
8	547.6	各種無線、無線局	250
9	547.7	放送無線	9
10	547.8	テレビジョン	165
11	547.9	高周波応用、超音波応用	1

2.2. 評価実験手順

あらかじめ用語辞書と同意語辞書を作成しておき、書誌データ（書名とNDC）を読み込み、以下の手順によって、階層型ニューラルネットワークに与えられる入力データおよび教師データを作成する。

① 同意語辞書を使用し、書名を修正する。
② 用語辞書を使用し、書名に使われている用語を抽出する。
③ 用語は出現順に従って、入力ユニット 1、2、3、．．．に割り振る。従って、書名は一般に複数の入力ユニット番号の組として表される。
④ NDCは1次分類と2次分類をそれぞれ4ビットで表現する。

ここで用いる同意語辞書は、例えば、「ファクシミリ」、「ファックス」、「Fax」、「FAX」を「FAX」に統一するためのもので、意味まで含めた同意語辞書ではない。また、それぞれ出現頻度が高い「計算機」と「コンピュータ」、「テレビ」と「TV」などは、同意語扱いとしない。出現頻度の高いものは、同意語扱いしても、学習効果の向上が期待できないためである。

Auto Classification of Books Using Neural Networks
Minoru Hatada
Systems Development Laboratory, Hitachi Ltd.
1099 Ohzenji, Asao-ku, Kawasaki, 215-0013 Japan

2.3. 評価実験結果

2.3.1. 学習曲線

評価実験結果を図1に示す。ここで、学習率1、認識率1は 1次分類、学習率2、認識率2は2次分類に対する値である。例えば、学習回数50回での認識率は、認識サンプルに50回の学習で得られた結合係数を用いて分類を求めたときの正解率を表わす。

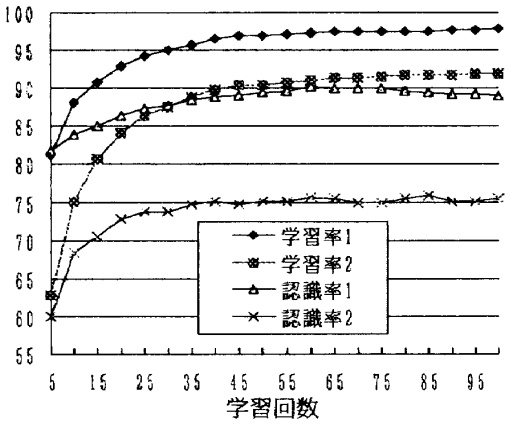

図 1. 学習曲線

2.3.2. 学習サンプル数と認識率の関係

学習サンプルを大量に入手するのは容易ではない。従って、なるべく少ない学習サンプルによって学習効果を得たい。そこで、学習サンプル数によって学習率、認識率がどう変わるかを調べた。

学習サンプル数と出現用語数の関係を図2 に示す。出現用語数には、500件の認識サンプルの書名に含まれる用語もカウントされている。

図2から、学習サンプル数が4,000を過ぎたあたりから、出現用語数にあまり顕著な伸びは見られない。このことから、認識率の向上も4,000サンプル辺りで伸び悩むことが予想される。

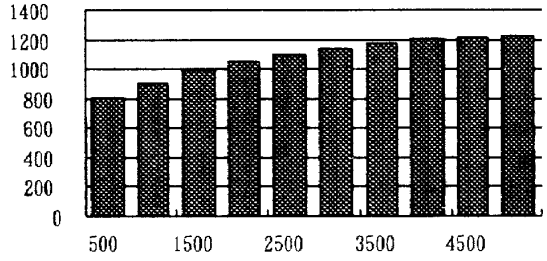

図2. 学習サンプル数と出現用語数の関係

図 3 に、学習サンプル数と学習率、認識率の関係を示す。

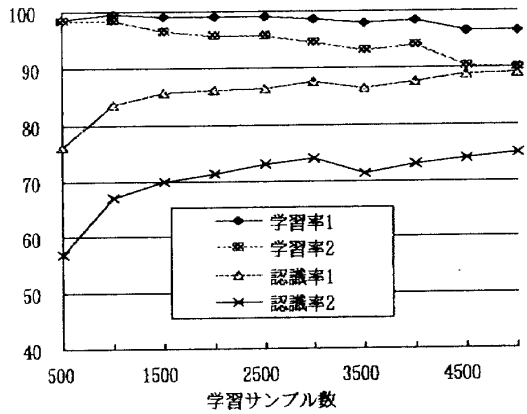

図3. 学習サンプル数と学習率･認識率の関係

どういうケースで認識エラーが発生するかを考察する。専門書の場合、本の内容を代表する技術用語を含んだ書名が付けられるが、書名から内容が判断できないケースが稀にある。例えば、「マジック・ボックス」、「日本の技術」、「驚きの材料」などである。今回のサンプルデータでは、こういうケースは 3～4% である。全ての用語を用語辞書に持てば、このようなケースは必ずしも学習エラーとはならないが、通常は認識に失敗する。

新しい用語で、学習サンプルにこれを含む書名が無い場合、当然、認識エラーとなる。例えば、「情報スーパーハイウェイ」、「遺伝アルゴリズム」などがこれに当たる。このようなケースは比較的少なく、せいぜい1%程度である。

3. おわりに

ニューラルネットワークによる図書の自動分類を行い、1 次分類では、認識率が最高 92%、2 次分類では最高 80%弱という結果を得た。1 次分類に対しては、まずまずの値と言える。しかし、蔵書数が数千冊を超えると、1 次分類では不足であり、2 次分類以上が必要となる。2 次分類での認識率 80%は、実用上充分な値ではなく、認識率改善に向けた研究が必要である。

参考文献

[1] 堤　一義：ニューラルネットワーク研究の最新動向、システム/制御/情報, Vol.36, No.10, pp.619-624 (1992)