人文科学とコンピュータ　37－6
（１９９８．１．３１）

年齢を表記した古文書文字の認識
－「宗門改帳」古文書画像データベースを用いた実験－

日置　慎治・上原　邦彦・川口　洋

帝塚山大学経済学部

〒６３１　奈良市帝塚山７－１－１

江戸時代に作成された「宗門改帳（しゅうもんあらためちょう）」を原史料として古文書画像データベースを構築する際，最も長い作業時間を必要とする過程は，史料読解，文字データ入力である。この作業過程を自動化するための基礎研究として，年齢を表記した 16 種類の古文書文字（一，二，三，四，五，六，七，八，九，十，壱，弐，拾，廿，ツ，年）を３層のニューラルネットを用いて認識する実験を行った。「宗門改帳」古文書画像データベースから 16 種類 1559 字を採字して，各 20 字を教師データとして学習させ，残りの未学習文字の認識率を求めた。実験の結果，認識率は 60～70％に達した。

Recognition of Hand Writing Historical Japanese Characters

Shinji HIOKI, Kunihiko UEHARA, Hiroshi KAWAGUCHI

Faculty of Economics, Tezukayama University

7-1-1 Tezukayama, Nara, 631, Japan

We tried to recognize 16 kinds of hand writing historical Japanese characters（一，二，三，四，五，六，七，八，九，十，壱，弐，拾，廿，ツ，年） by the use of neural network with 3 layers. We took 1559 characters from the image database of the Japanese religious investigation registers as test data. Also, we chose 20 characters per each kind from the test data as the supervised data. Then we tested 1239 unlearning characters by this neural network system. The recognition rate reached 60-70%.

1 はじめに

1．1 研究の目的

筆者は，江戸時代に記録された古文書史料から人口学的指標を計算する作業過程をできる限り自動化するシステムの構築を計画している。このシステムを「江戸時代における人口分析システム」と呼び，略称を DANJURO（Demographic analyzing system in Tokugawa Japan using the relational database of Shumon-Aratame-Cho）と名付けた。DANJURO ver.1 は，史料を読解した文字データを登録したデータベースと人口学的指標を出力するプログラム群から構成されていた[1)]。手作業で行われていた従来の研究方法と比較すると，データベースに登録された文字データから人口学的指標を算出する作業時間は大幅に短縮された。ver.1 では，作業時間の一層の短縮，研究過程に関する再現性の確保，史料の保存，研究者間における史料と分析方法の共有といった点が未解決のまま残されていた。これらの課題を改善して ver.2 を開発するための基礎研究として，古文書画像データベースの構築を開始した[2)]。

古文書画像データベースを構築する場合，最も長時間の作業を必要とするのが，史料読解，文字データ入力である。古文書読解技能を持つ研究者の手作業によって行われてきた史料読解，入力作業を部分的にでも自動化できれば，史料読解から人口学的指標算出までの作業時間を飛躍的に短縮することができる。そこで古文書文字の自動認識に関する実験を開始した。

1．2 史料の概要

江戸時代の日本では，「宗門改帳（しゅうもんあらためちょう）」と総称される古文書史料が，17 世紀末から 19 世紀中期の明治初年まで全国で作られていた。たとえば，陸奥国会津郡，大沼郡，下野国塩谷郡（現在の福島県南会津郡，大沼郡，栃木県塩谷郡）の一部を含む南山御蔵入領（みなみやまおくらいりりょう）では，元禄 7（1694）年あるいは元禄 8 年から明治 3（1870）年まで，図 1 のような書式の史料が，毎年，村ごとに名主の手によって作成され，代官所と名主の自宅に 1 部づつ保管されていた。南山御蔵入領に所属する小松川村（現在，福島県南会津郡下郷町）には，散逸した 8 年分を除いて，寛政 4（1792）年から慶応 2（1866）年に至る 75 年間の「宗門改人別家別書上帳」が保存されている。この史料には，記載単位ごとに，旦那寺の本末関係，所在地，宗派，旦那寺の名称，持高，家屋規模，屋根の材料，構成員の名前，筆頭者との続柄，年齢，異動，牛馬数，世帯規模などが記録されている。「宗門改帳」の全国的な所在調査は進行中だが，19 世紀中期の日本に存在した 63562 ヶ村の約 3 %，1900 余りの村において，良質の史料が保存されている可能性があると思われる。

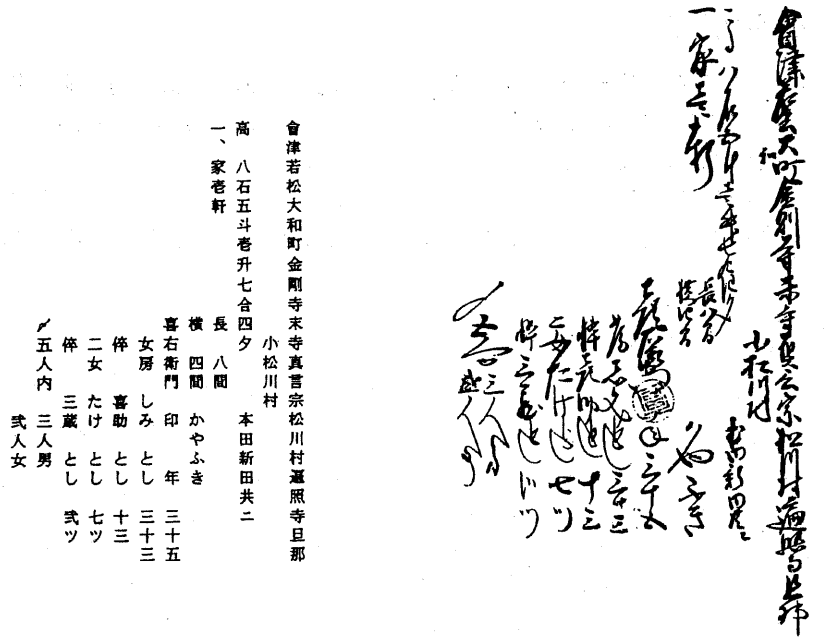
図 1 のような史料が継年的に保存されている村では，人口変動のほかにも，初婚年齢，死亡年齢，養子や婚姻による人口移動の範囲といった人口再生産構造に影響を持つ人口学的指標，家族形態，家族周期，相続や改名に関する慣習など，民衆生活の具体像を示す情報を求めることができる。

1．3 「宗門改帳」古文書画像データベースの概要

小松川村の「宗門改人別家別書上帳」を入力史料として，「宗門改帳」古文書画像データベースを試作した。本データベースは，ORACLE ver.8.0 を DBMS として，ア）個人情報，イ）世帯情報，ウ）

古文書画像情報，エ）史料書誌情報という４つのテーブルから構成されている。

文字認識の実験対象となる文字を採字したウ）に登録されている古文書画像は，次の手順でデジタル化されている。まず，「宗門改人別家別書上帳」の見開き２ページを１画像として写真撮影して，PHOTO－CD を作成した。次に，PHOTO-CD から 1536＊1024 DOTS の解像度で読み込み，256 階調のグレイスケールに調整，フィルター（シャープ強）をかけ１世帯を１画像に編集後，JPEG 形式で保存した。



会津若松大和町金剛寺末寺真言宗松川村遍照寺旦那
小松川村
高　八石五斗壱升七合四夕　本田新田共ニ
一、家壱軒
長　八間
横　四間　かやふき
喜右衛門　印　年　三十五
女房　しみ　とし　三十三
倅　喜助　とし　十三
二女　たけ　とし　七ツ
倅　三蔵　とし　弐ツ
五人内　三人男
弐人女

図１　「宗門改帳」の書式

史料：「文久三年　宗門家別人別改書上帳　小松川村，沢入村，寺山村，寺村」（佐藤仁夫家所蔵）

2 古文書文字の採字

2. 1 実験対象とした古文書文字

年齢を表記した「一，二，三，四，五，六，七，八，九，十，壱，弐，拾，廿，ツ，年」の 16 種類の文字を対象として，古文書文字認識の実験を開始した。

「文政八年酉年二月　宗旨家別人別分限書上帳　小松川村，寺山村，大久保村，沢入村，寺村」のうち小松川村分の史料には，273 種類，3,898 文字が使われている。このうち今回の実験対象として選択した 16 種類の文字は，全体の約 22％に相当する 868 文字出現する。とくに，「弐」，「壱」，「四」は，出現順位が 10 位以内に入る頻出文字である。加齢などにともない史料作成年次ごとの出現順位は変化するが，16 種類の文字は常に頻出する。

漢数字で表記される年齢，牛馬数，世帯規模，持高，家屋規模といった情報の中で，年齢は，結婚年齢，出産年齢，死亡年齢，夫妻の年齢差，年齢別人口構成，生命表といった人口学的指標を算出する場合，とくに重要な基礎的情報となる。年齢を表記した漢数字の種類は限定されるうえに，図１のように世帯構成員の名前の下のほぼ固定した位置に記録されているため，セグメンテーションも比較的容易と

予測される。

本研究の対象となる古文書文字は，和紙に毛筆で書かれた手書き文字である。①1種類の文字であっても，字形，字体に相当のばらつきが見られる，②続け字が多用されている，③文字の太さが多様である，④前後の文字などの影響で，文字の大きさが多様であるといった特徴を持っている。

2. 2 採字の方法

「宗門改帳」古文書画像データベースに登録されている古文書画像から，実験対象文字のうち「廿」を除いた15種類の文字を各100個づつ採字した。「廿」については，59個しか採字できなかった。したがって，実験対象は16種類，1559個の古文書文字である。

採字は，①PAINT SHOPを用いてグレイスケールの原画像を2値化する，②古文書文字を切り出す，③ビットマップ形式に変換する，④1文字を1ファイルに保存するという手順で実行した。

3 古文書文字認識の方法

古文書文字認識の方法としては，テンプレート・マッチングや特徴抽出など様々な提案がされている[3]。本研究では，全般的な方法として，ニューラルネットを用いた手書き文字認識と同様の方法を採用した[4]。この方法を選択した理由は，特徴抽出などによる識別処理に比べ，文字の持っているすべての特徴を取り入れられると期待できることによる。もちろん，認識率に寄与する顕著な特徴が抽出されれば，入力層のPE（プロッセッシング・エレメント）に加える方針である[5]。

ニューラルネットによる古文書文字認識の処理は，次の4段階に大別される。

① セグメンテーション：レイアウト解析（劣化画像復原を含む），行切り出し，文字切り出し
② 濃度計算：標本点選択，マスクパターン決定，濃度特徴計算
③ ネットワーク出力計算：学習済み多層ニューラルネットによる出力PE計算
④ 最大出力PE検出：最大出力PE検出，対応文字決定

①については先行研究[6]を活用することとして，ここでは②以下について実験を行った。

3. 1 濃度特徴計算

濃度特徴計算では，縦横比，大きさの多様な入力画像をニューラルネットの入力層のデータとするため，縦Z点横Z点の多値データとして正規化した。まず，入力画像の縦横のうち小さい方の両側に空白を付け足して正方形にした。次いで，標本点の近傍の点に重みをつけながら取り入れるマスクをかけ，Z＊Zに正規化する。この際，文字の重心がZ＊Zの中心に来るように平行移動させた。

3. 2 ネットワーク出力計算

ニューラルネットの学習過程では，濃度値を入力層からの出力としてバックプロパゲーション法で学習する。これを入力層，中間層および出力層の3層からなるニューラルネットの範疇で実行し，結合係数を出力層各PEの教師付き学習により求める。

0	0	0	0	20	60	20	0	0	0
0	0	0	120	180	180	40	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	40	180	140	0	0	0	0	0
0	0	140	140	40	0	0	0	0	0
0	120	100	0	0	20	120	60	100	60
120	120	80	120	160	180	180	180	180	160
180	180	180	180	180	160	120	120	120	80
0	0	0	0	0	0	0	0	0	0

（三）の場合

0	0	0	80	160	40	0	0	0	0
0	0	100	180	180	60	0	0	0	0
20	0	0	140	180	60	0	0	0	0
40	100	160	180	120	20	140	120	160	100
20	80	0	180	180	120	20	0	120	180
40	160	0	180	140	20	0	0	160	140
120	80	40	180	60	0	0	80	180	20
120	60	0	180	60	0	20	180	20	0
120	60	0	80	20	0	120	80	0	0
120	40	0	0	0	40	140	0	0	0

（弐）の場合

図2　濃度特徴計算の結果

入力層は，濃度特徴計算の結果正規化されたＺ＊Ｚの多値データである。Ｚを 10 とした場合には，100 点多値データとなる。中間層のＰＥ数は 30 を採用した。中間層ＰＥ数の適正規模については，今後調節する予定である。出力層については，16 種類の文字認識を前提としていたので，16 個の状態を区別できる４ビットという意味で４とした。16 の状態を区別しているという意味で出力層は 16 と表現する場合もあるが，プログラム上ではこのように簡単化している。

教師付き学習の過程は，ア）～カ）の表記を用いて７ステップにまとめることができる。

ア）入力層ニューロン……U1(k,p)

k は入力データの点を表す指標である。Ｚ＊Ｚに正規化した場合には，k=1,2,3,...,Z*Z となる。Ｚ＝10 の場合には k=1,..,100 である。p は教師データを区別する指標で，16 種類の文字，各Ｎ個を学習させる場合には，p=1,2,3,...,16＊Ｎ となる。

イ）中間層ニューロン……U2(i,p)

i は中間層 i 番目のニューロンの状態を表す指標である。中間層が 30 の場合には，i=1,2,3,...,30 となる。p はア）と同様，教師データを区別する指標である。

ウ）出力層ニューロン……U3(j,p)

j は出力層 j 番目のニューロンの状態を表す指標である。出力層が 16 の場合には，i=1,2,3,...,16 となる。p はア）と同様，教師データを区別する指標である。

エ）入力・中間層結合……C12(i,k)

k,i はそれぞれア），イ）で定義された入力層，中間層のニューロンの番号を表す。

オ）中間・出力層結合……C23(j,i)

i,j はそれぞれイ），ウ）で定義された中間層，出力層のニューロンの番号を表す。

カ）正解データ……T(p)

教師データ p に対する正解を表す。

step1）　C12，C23 を乱数で初期化する。

step2）　ある教師データ p に対して U1(k,p)を用意する。

step3）　次式にもとづいて U2 を求める。

$U2(i,p)=f(\Sigma_i\{C12(i,k)*U1(k,p)\})$

関数 f(x)は，$f(x)=(1+\tanh(x))/2$ と定義されるシグモイド関数を用いた。

step4) U3 を求める。

$U3(j,p)=f(\Sigma_k\{C23(j,i)*U2(i,p)\})$

step5) p に対する正解 T(p)と U3 との差を利用して，評価関数 E を次式で定義する。

$E=\Sigma_{pj}\{(T(p)-U3(j,p))^2\}$

step6) 評価関数 E を小さくする方向に，C12,C23 を以下のように少し修正する。

$C23(j,i)=C23(j,i)+e*\Sigma_p\{(U3(j,p)-T(p))\,f'(\Sigma_n\{C23(j,n)*U2(n,p)\})\}*U2(i,p)\}$

$C12(i,k)=C12(i,k)+e*\Sigma_p\{f'(\Sigma_n\{C12(i,n)*U1(n,p)\})\ *$

$\Sigma_j\{C23(j,i)*\Sigma_p\{(U3(j,p)-T(p))\,f'(\Sigma_n\{C23(j,n)*U2(n,p)\})\}*\ U1(k,p)\}$

e は予備実験から 0.4 とした。

step7) E がある閾値（例えば 0.01 など）よりも小さくなるまで step2）～step6）を繰り返す。この繰り返しの数を iteration 数と呼ぶ。

4 実験結果

採字した 16 種類，1559 文字から各 20 文字を教師データとして学習させ，残りの未学習文字の認識率を求めた。A，Bの異なる2セットの教師データについて実験を行い，Bセットについては，入力層を 10＊10 にした場合と 16＊16 にした場合を比較した。使用した CPU は DEC の Alpha 21164 である。学習に要した計算時間は5分から 60 分である。学習は 5000 から 50000 回の iteration で収束した。

結合係数を乱数で初期化するため，教師データのすべてを正解する結合係数が多数存在する場合，学習後の結合係数は初期値に依存する。教師データを増やすと結合係数の空間が縮小し，初期値依存性は減少すると予測される。しかし，1文字当たり 40 個の教師データを使った時にも初期値依存性が確認された。今回は 10 セットの乱数列を用いて実験を行い，誤差を求めた。

Aセットについて入力層を 10＊10 として 10 セットの異なる乱数列で計算した結果，未学習文字 1239 のなかで正しく認識された文字数の平均は，770+/－ 9（+／－9は初期値による誤差である。以下このように表記する。）となった。最も認識率の高かった場合の各文字の認識率は，表1－1に示される。認識率は 16 文字全体で 65％となるが，「一」，「ツ」，「年」は 90％を越えるのに対して，「五」，「七」，「九」，「弐」，「拾」は 50％以下である。

Bセットについて入力層を 10＊10 として 10 セットの異なる乱数列で計算した結果，未学習文字 1239 のなかで正しく認識された文字数の平均は，860+/－ 12 となった。最も認識率の高かった場合の各文字の認識率は，表1－2に示される。認識率は 16 文字全体で 75％となるが，「一」，「三」，「年」は 90％を越えるのに対して，「八」，「弐」は 60％以下である。

Bセットについて入力層を 16＊16 として 10 セットの異なる乱数列で計算した結果，未学習文字 1239 のなかで正しく認識された文字数の平均は，840+/－ 8 となった。最も認識率の高かった場合の各文字の認識率は，表1－3に示される。認識率は 16 文字全体で 68％となるが，「一」，「年」は 90％を越えるのに対して，「五」，「六」，「九」，「壱」は 50％以下である。

表1　古文書文字の認識率

1－1　学習文字Aセットの場合(入力層:10＊10)

文字の種類	一	二	三	四	五	六	七	八	九	十	壱	弐	拾	廿	ツ	年	合計
未学習文字数	80	80	80	80	80	80	80	80	80	80	80	80	80	39	80	80	1,239
正しく認識された文字数	74	43	66	47	40	54	32	58	31	55	54	36	38	29	73	79	809
認識率(%)	93	54	83	59	50	68	40	73	39	69	68	45	48	74	91	99	65

10セットの異なる乱数列での計算結果:正しく認識された文字数の平均(認識率)=770(62%)+/- 9

1－2　学習文字Bセットの場合(入力層:10＊10)

文字の種類	一	二	三	四	五	六	七	八	九	十	壱	弐	拾	廿	ツ	年	合計
未学習文字数	80	80	80	80	80	80	80	80	80	80	80	80	80	39	80	80	1,239
正しく認識された文字数	77	67	73	55	53	54	55	44	61	55	58	48	58	30	71	73	932
認識率(%)	96	84	91	69	66	68	69	55	76	69	73	60	73	77	89	91	75

10セットの異なる乱数列での計算結果:正しく認識された文字数の平均(認識率)=860(69%)+/- 12

1－3　学習文字Bセットの場合(入力層:16＊16)

文字の種類	一	二	三	四	五	六	七	八	九	十	壱	弐	拾	廿	ツ	年	合計
未学習文字数	80	80	80	80	80	80	80	80	80	80	80	80	80	39	80	80	1,239
正しく認識された文字数	76	70	71	48	36	39	46	58	38	46	36	56	50	31	66	73	840
認識率(%)	95	88	89	60	45	49	58	73	48	58	45	70	63	79	83	91	68

8セットの異なる乱数列での計算結果:正しく認識された文字数の平均(認識率)=801(65%)+/- 8

5　おわりに

本稿では，「宗門改帳」古文書画像データベースを構築する際，最も長い作業時間を必要とする史料読解，文字データ入力の過程を自動化するための基礎研究として，年齢表記に用いられる 16 種類の古文書文字をニューラルネットを用いて認識する実験について報告した。認識率は60～70%にのぼった。

認識率を上げるためには，濃度特徴計算を行う前に，細線化を含む前処理を試みるとともに，多様な形状をしている古文書文字の正規化の方法を確立する必要がある。入力層，中間層の PE 数についても，今後調節したい。さらに，字形の種類に対応させてネットワークを多層化することも視野に入れる必要があろう。

学習文字の数を増加させると，学習が収束しなくなる可能性がある。学習がローカルミニマムに陥るのを回避する多層間の結合係数の最適化法として，シミュレーテッド・アニーリング（疑似焼きなまし）法を用いたい。この方法は多大な計算時間を必要とするので，計算の高速化法が必要となる。高速化法としては，巡回セールスマン問題（TSP）に適用され，ある程度の正答率が得られる実空間繰り込み群的アプローチなどがある。

付記)

本報告は，平成９年度文部省科学研究費補助金（重点領域研究）「人文科学とコンピュータ」公募研究（課題番号：09204236），平成９年度日本私学振興財団学術研究振興資金，および平成９年度帝塚山大学特別研究費の補助を得て行った研究成果の一部である。

注）

1）DANJURO ver.1 については以下の文献に詳しい。
川口　洋（1990）江戸時代における人口分析の方法　－奥会津地域における「宗門改人別家別帳」のデータベース化を事例として－，歴史地理学，no.151，pp.16-33
川口　洋（1993）「宗門改帳」データベース・システム（DANJURO）の改良，情報処理学会研究報告「人文科学とコンピュータ」，vol.92,no.19，pp.1-8
川口　洋（1995）コンピュータを用いた江戸時代における人口分析の方法，人文学と情報処理，no.7，pp.54-58

2）「宗門改帳」古文書画像データベースについては以下の文献に詳しい。
川口　洋・上原邦彦（1996）「宗門改帳」を入力史料とした古文書画像データベースの構築，情報処理学会研究報告「人文科学とコンピュータ」，vol.96,no.110，pp.49-54
川口　洋・上原邦彦（1997）古文書画像データベースの構築　―「宗門改帳（しゅうもんあらためちょう）」を入力史料として―，人文学と情報処理，no.14，pp.49-55

3）以下の文献に古文書文字認識に関する試みが報告されている。
山田奨治（1995）高次局所自己相関特徴による古文書かな文字認識，情報処理学会研究報告「人文科学とコンピュータ」，vol.95,no.14，pp.21-30
柴山　守他（1997）古文書画像の文字セグメンテーションとツール開発，京都大学大型計算機センター第57回研究セミナー報告，pp.3-9

4）小川英光編著（1994）『パターン認識・理解の新たな展開』電子情報通信学会，pp.36-41
国際電気通信基礎技術研究所編（1995）『ニューラルネットワーク応用』オーム社，pp.27-91

5）文字の特徴をうまく取り入れることにより，例えば「一」に関しては，上から下に向かって白黒反転が一回であるという条件で，他の文字と区別することができよう。ただし，特徴抽出による文字認識は，実験対象となる文字の種類が増えると，人間に依存する作業の重要性が大きくなり，対応が困難になる可能性が危惧される。

6）セグメンテーションに関する研究例として，次の文献があげられる。
富田浩章・柴山　守・荒木義彦（1996）2値化レベル制御による古文書画像の文字セグメンテーションとパターン字書について，情報処理学会研究報告「人文科学とコンピュータ」，vol.96 no.42，pp.7-12