RJ-006

人間乱数による個人識別の可能性

Possibility of Personal Identification by means of Human Random Generation

田中 侑希†　　田中 美栄子†
Yuuki Tanaka　　Mieko Tanaka-Yamawaki

1. はじめに

近年，コンピュータや携帯電話の普及により身近に個人認証システムを利用する場面が増えて来た．それに伴い，有効で使いやすい個人認証技術への期待が高まっている．本稿では人間乱数による個人認証の可能性について考察する．現在用いられている個人認証技術には知識属性（パスワード方式），所有物属性（クレジットカード，IC カード等），生体属性（指紋，網膜，声紋等）を用いたものが存在するが，それぞれ順に，知識の忘却，所有物の紛失，生体情報登録への抵抗や登録情報の変更不可という欠点を持つ．そこでこれらの問題点を補う方法として人間乱数を利用することを考えた．人間乱数を利用した認証システムを作ることができれば，暗記や物の所持の必要がなく，生体情報登録への心配もない，理想的な個人認証システムとなり得ると考えられる．

2. 研究目的

人に「デタラメになるように数字を並べてください」と言う作業のことを人間乱数テストと呼び，作成された数列を人間乱数（列）と呼ぶ．得られた数列は完全な乱数列ではなく生成者の癖や状態が反映された特徴的な数列になることが知られている[1] ．また，同じ人が生成した数列でも生成方法により乱数の性質が変化することも分かっている[2] [3] [4] [5] ．そこで人間乱数を個人認証に利用するために，個人の癖（特徴）が数列に強く反映されるような生成方法を探ることにした．人の癖は直感的な行動を取ったときに現れると考えられることから，乱数生成の時間を制限することで，熟考ではなく直感的に生成させ，それによって生成された数列にその人特有の癖が反映されると考えた．本研究では 2 秒以下の短時間の思考を直感と定義する．

本研究は，直感的に生成した乱数を様々な指標を用いて解析することによって数列の特徴を捉え，また SOM を用いてデータの個人分類可能性を検証することで，人間乱数による個人識別の可能性を示すことを目的とする．

3. 乱数生成法

個人認証システムへの応用という目的に従い，数列に個人の癖が反映される方法を探るため，乱数を直感的に生成する方法を提案する．提案採取手法についての条件をまとめたものが表 1 である．使用するシンボルは一般的に人間乱数テストで用いられる 0～9 の整数，乱数列の長さは 50，データの採取にはコンピュータにテンキーを接続したものを記録装置として使用する．乱数列の生成は直前に入力した数が不可視となるプログラムを使用して被験者自身が入力を行う．実際の生成画面を図 1 に示す．直前の入力を不可視にした理由はデータに脳の状態を反映させるために，視覚情報ではなく脳内の記憶を利用してもらうためである．プログラムは最初にユーザ名を入力し，その後被験者の任意のタイミングで入力を開始し，乱数を 50 個入力した所で自動的に決められた書式のファイルを出力するものを作成した．

また，直感的と判断する目安として入力間隔を 2 秒以内とするが，その根拠としては五味等の報告[3] より，生成速度が 5 秒間隔以上と 1 秒間隔以下で生成したデータに大きな違いが見られたことから，1 秒間隔以下を直感的に生成したと判断する一応の目安にした．しかし 1 秒間隔以下の入力は時間を意識してしまい焦りから正しく被験者のデータが生成されないと考え，時間を意識しなくても生成できる入力間隔として 2 秒間隔以内とした．但し思考度合が重要であるため時間については気にしないよう被験者に強く説明した．

さらに，テスト開始前の被験者への説明は簡単にプログラムの使い方を説明した後に「深く考えず直感的に乱数になるように数を入力してください」と指示した．直感的の意味はは短時間，具体的には 2 秒以内と説明し，乱数がどのようなものかについては個人で考えてもらうようにした．これは乱数がどのようなものかを説明するのが困難であり，説明の仕方によって被験者に誤った認識を与えてしまうことを防ぐためである．こうして得られたデータは長さ 50 のデータ生成を 1 試行とし，1 試行 1 ファイルとして扱う．

表 1 提案採取方法の条件一覧

条件	パラメータ
シンボル	0~9 の整数（10 種類）
乱数列の長さ	50
生成方法	直前の数が不可視となるプログラム
記録方法	テンキーを接続した PC
記録者	被験者
制約条件	直感的生成（2 秒以内の間隔）

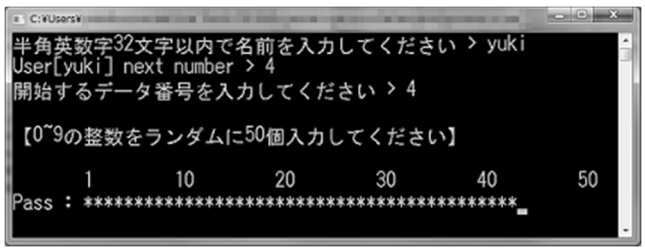

図 1 乱数生成プログラム実行画面

本研究で使用したデータは，同大学の工学部情報系学科に所属する 21～23 歳の男性 6 名（A～F）から一人当たり 50 ファイル採取した．

また，データ採取は一度に最大 10 試行までとし，採取日時は各々の判断で比較的気持ちが落ち着いているときに

†鳥取大学大学院工学研究科情報エレクトロニクス専攻

被験者のみで行った．ここで直感的に考えて生成した乱数列を直感人間乱数と呼ぶことにする．また乱数の比較用に擬似乱数として rand() 等で用いられる線形合同法（LCG）のデータ（Seed:1～50）も用意した．

4. 乱数列の特徴を捉える指標

乱数列から特徴を捉える際ただ数列を眺めてもよくわからない．そもそも乱数とは本来規則性のないものである．そこで乱数列を比較するために数列の規則性を数値化できる指標を用いて乱数列の特徴を捉える．今回解析に使用した指標は 8 種類である[5] [6] [7]．これは，個人の癖がどのように現れるか分からないため，多面的に特徴を探ろうと思い 8 種類の指標を用いた．各指標については 4.1～4.8 節で説明する．また，各被験者の直感人間乱数に擬似乱数を加えた 7 者に対して各指標の平均値をまとめたものを表 2 に示す．表 2 は直感人間乱数 50 ファイルの平均値を示す．また，括弧内は標準偏差を示している．

表 2 直感人間乱数の被験者 6 名（A～F）と LCG（L）の各指標（H,…,TIME）の平均値（標準偏差）

	H	ADJ	TPI	PL
A	.978(.017)	7.4(4.5)	114.3(8.7)	.16(.54)
B	.991(.004)	14.5(5.1)	106.3(10.9)	.36(.86)
C	.983(.010)	24.9(6.8)	86.8(10.5)	1.97(1.72)
D	.970(.016)	23.1(6.2)	89.9(10.6)	1.18(.86)
E	.949(.028)	26.1(8.3)	87.3(12.8)	1.38(1.98)
F	.978(.011)	26.8(6.4)	94.6(9.7)	1.46(1.62)
L	.959(.019)	17.2(5.1)	96.5(8.5)	.83(1.25)
	RG(mean)	RP	TKD	TIME
A	9.20(.58)	47.4(10.4)	56.0(2.5)	.54(.07)
B	9.77(.17)	44.6(8.2)	51.9(4.0)	.97(.22)
C	9.39(.30)	40.4(6.9)	46.1(3.7)	.81(.20)
D	8.89(.38)	48.2(8.2)	50.3(3.8)	.33(.06)
E	8.77(.55)	52.3(9.7)	50.9(6.1)	.44(.13)
F	9.27(.42)	45.7(9.5)	44.9(3.1)	.49(.04)
L	7.79(.65)	36.5(6.8)	48.2(3.7)	

4.1 Entropy（H）

Entorpy はシンボル生成偏差を表し，次式で定義される．これは，シャノンエントロピーと同様のものである．

$$H = -\sum_i P_i \log P_i \qquad (1)$$

この時，P_i はi 番目のパターンの出現確率を示す．ここでは$i = 0$~9 の値を取り，一桁の数の生成偏差を求めている．H は 0～1 の値を取り 1 に近い時シンボルの生成偏差が均一であることを示し，逆に 0 に近いほど偏った生成偏差を示す．一様乱数である場合，有限の区間で区切られた区間内の実数は同じ確率で現れる．つまり 1 に近いほど乱数度が高いと言える．

4.2 Adjacency（ADJ）

Adjacency とは隣接する文字の演算差の絶対値 1 が出現する確率を表すものであり，次式で定義する．

$$ADJ = 100 \times \frac{NAP}{m-1} \qquad (2)$$

m は数列の長さを表しここでは$m = 50$である．NAP は隣接する数字の絶対値 1 の個数を示す．ADJ は 0～100 の値を取り 0 に近いほど出現確率が低いことを示し，100 に近いほど出現確率が高いことを示す．

4.3 Turning Point Index（TPI）

Turning Point Index とは数列の上昇と下降が切り替わるポイント Turning Point(TP）の出現回数を式(3)の期待値と比較した値で定義される．

$$TP_{ex.} = \frac{2}{3}(m-2) \qquad (3)$$

$$TPI = 100 \times \frac{TP_{ob.}}{TP_{ex.}} \qquad (4)$$

$TP_{ex.}$は実際に測定した TP の出現回数，$TP_{ob.}$は乱数の TP 出現回数の期待値を示す．すなわち TP の出現回数が多いと 100 以上の値になり逆に少ないと 100 以下の値となる．ここで例として数列「5,3,4,6,2,8,9,7,1,0」があったとすると TP は「3」「6」「2」「9」と 4 回出現したので$TP_{ob.} = 4$ となる．

4.4 Phase Length（PL）

Phase Length とは TP が発生する間隔距離d の出現回数の期待値に対する比で表す．

$$PL_{ex.}(d) = \frac{2(m-d-2)(d^2+3d+1)}{(d+3)!} \qquad (5)$$

$$PL(d) = \frac{PL_{ob.}(d)}{PL_{ex.}(d)} \qquad (6)$$

ここで$PL_{ob.}(d)$ は実際に測定した$PL(d)$ の出現回数，$PL_{ex.}(d)$ は$PL(d)$ の期待値を示す． d は$[1, m-3]$ の範囲の値を取る．

例えば数列「2,3,5,4,5,6,7,8,6,1,3」に対して， TP は「5」「4」「8」「1」の 4 点であり，各 TP 間の長さは「1」「4」「2」となるため，$PL_{ob.}(d = 1,2,4) = 1$，$PL_{ob.}(d = 0,3,5,..) = 0$ となる．

各d の期待値を$m = 50$ の場合に求めたものが表 3 である．これをみると$d = 5$ から期待値が大幅に減少する．これは数列の長さによる影響でm に比例してd の期待値も増減する．今回$d = 5$~47 は使用しないこととした．また$d = 1,2,3$ についてもあまり有効な特徴が見られなかったためここでは$d = 4$ を使用する．

表 3 *PL (d)* の期待値（*m*=50）

d	期待値	d	期待値
1	19.583333	6	0.012731
2	8.433300	7	0.001604
3	2.375000	8	0.000178
4	0.506349	9	0.000018
5	0.087450	10~47	0.000002

4.5 Repetition Gap（RG）

Repetition Gap とは同じシンボルが繰り返し出現する距離 Repetition Distance(RD)の値を解析する指標である．今回は RD の平均値（mean）を使用する．ここで例として数列「2,3,7,8,8,7,2,3,2」があったとするとシンボル「2」は 6 個先とさらにその 2 個先に出現する．シンボル「3」は 6 個先にシンボル「7」は 3 個先，シンボル「8」は 1 個先に出現する．これをまとめると表 4 のようになる．この表から繰り返し距離の平均値を求めた値が RG である．

表 4　RD の出現回数

繰り返し距離 d	1	2	3	4	5	6	7	8
出現回数	1	1	1	0	0	2	0	0

4.6 Repeat Pattern（RP）

Repeat Pattern とはデータ中における隣接する 2 文字の繰り返し出現する頻度を表したもので次式により定義される．

$$RP = \left(1 - \frac{NRS}{m-1}\right) \qquad (7)$$

m は数列の長さを表しここでは$m = 50$ である．NRS は一度しか出現しなかったパターンの個数を示す．RP は 0～100 の値を取り，0 に近いほど繰り返しのパターンの生成が多いことを示す．

4.7 Ten Key Distance（TKD）

Ten Key Distance とは数を入力する際のキーの総移動距離を表したものである．キー間の距離を図 2 のような座標に置き換え入力したキーとその前に押されたキーとの距離を計算し合計したものである．0 については 2 つ分のスペースがあるため前後のキーとの最短距離になるように定めた．例えば数列「2,4,5,9,1,7,0,6」があったとき，それぞれの移動距離は$\sqrt{2}$,1, $\sqrt{2}$, 2$\sqrt{2}$,2,3, $\sqrt{3}$となる．

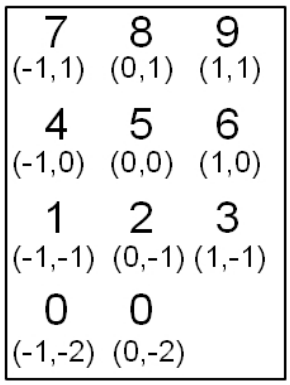

図 2　テンキーの座標

4.8 生成時間（TIME）

乱数を生成する際に生成間隔も同時に記録した．このデータを使い生成間隔の平均値を求めた．

5. 個人の癖の存在

直感人間乱数に個人の癖がどのように反映されているか表 2 を用いて指標ごとに検証する．

H は B，E のよう差異がある人もいるが全体的に個人間に明確な差は現れなかった．しかし，擬似乱数に比べると全体的に高い値を示しており，直感人間乱数と擬似乱数の区別に利用できそうである．

ADJ は今回使用した乱数指標の中で一番差異が見られた．これは，無意識の内に 0-1 や 7-6 といった連続した数字を出さないようにした人と，そうでない人で大きく分かれたと考えられる．つまり，連続した数を発生させないという意識の度合いが癖として現れていると言える．

TPI は ADJ と同じく差異が見られた．これは，数の大きさを意識して大小にした人と，そうでない人で大きく別れたと考えられる．つまり，数の昇順降順に対する意識の度合いが癖として現れていると言える．また，TPI と ADJ は r=-.73 と相関が強いことも分かっている．

PL は平均値を見ると差異はあるが個人内の差（標準偏差）も大きく不安定な指標であった．これは，本研究では長さ 50 の数列に対して d=4 を使用している．さらに期待値が 0.5 からも分かるようにちょっとした差であっても標準偏差が大きくなる傾向にある．しかし，d=4 を発生させる人としない人では明確な差がある．

RG は個人間の差は小さいが標準偏差も小さく個人内の差が小さいことが分かる．これは，無意識の内に同じ数を使うまでの長さを決めていると考えられる．また，人間乱数ではシンボルの数と同じ値になりやすい性質も見られる．なので，H と同様に直感人間乱数と擬似乱数の区別に利用できそうである．

RP は 個人間に差異が見られた．また，全体的に擬似乱数より高い値を取っている．これは，直感人間乱数には連続して発生させる数に癖が現れやすいと考えられる．よって，繰り返しの出現頻度ではなく発生した連続パターンを解析することで，より個人の癖が見られる可能性がある．

TKD，TIME は両方個人間かに差異が見られた．特に TIME は個人内の差が小さく個人間の差が大きかった．これは，乱数を発生させる際のリズムが癖として現れていると言える．この 2 つの指標は乱数を解析したものではなく，乱数発生時の人の動作を数値化したものであり，乱数の指標に比べて個人の癖が現れやすい傾向にあった．

以上のことより指標による差はあるが，直感人間乱数に個人の癖が存在しているのが分かる．また，直感でない人間乱数に対しても同様の解析をしたところ，乱数指標は ADJ を除いて大きな変化はなかったが，動作指標は個人内の差が大きく直感とは反対にとても不安定な指標となった．

6. SOM を利用した人間乱数の分類

単一の指標で個人を分類することはできないのが表 2 から分かる．そこで，前述の 8 個の指標から得られる特徴量を全て同時に用い，8 次元のデータとして扱うことで個人分類を試みた． 8 次元データの分類方法として，全指標が互いに独立とは限らない場合にも応用でき，また結果を分かりやすく可視化できる利点を持つ SOM を利用する．

6.1 SOM とは

自己組織化マップ（SOM：Self-Organizing Maps）はニューラルネットワークの一種で，多変量の特徴をもつデータ群を性質の類似度によって分類する教師なしクラスタリングである．

6.2 直感人間乱数の個人分類実験

直感人間乱数がそれぞれ個人で分類できるか実験を行った．直感人間乱数を一人各 40 ファイル使用して学習したマップを図 3 に示す．また，学習パラメータは量子化誤差が小さくなるように実験的に求めたものを使用した．

図 3 を見ると個人毎にデータが集まって配置されている．特に A，B は同一人物のデータが固まって配置されており他のデータと分類されている．C～F についても概ね同一人物のデータ毎に固まって配置されている．つまり，直感人間乱数データは個人毎に差異があることが分かる．よって直感人間乱数は個人識別の可能性があると言える．しかし，データ間の距離を表すマップの色が全体的に薄くデータ間の距離が近いことが分かる．特に B と C，C と F，D と E のように近くにあるデータは一部が混ざり合うように分類されているのが分かる．

次に図 3 上のマップが正しく学習されているかを調べるため個人識別実験を行った．使用したデータは学習で使用しなかった直感人間乱数一人各 10 ファイルを用いた．識別用のデータを図 3 の学習マップに入力した結果を図 4 に示す．図 4 を見ると A，D，E，F は学習マップと比べ同一人物のファイルと同じ場所に分類されている．しかし，C は近い場所に分類されているが多くのファイルが他のファイルと混ざり合う境界付近に分類されている．つまり，個人毎にデータの差異はあるがその差は小さく，明確に分類するこができていない．

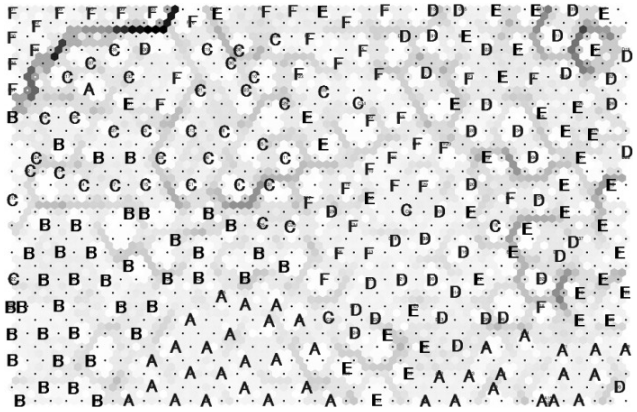

図 3　学習用データ 240 ファイルの分類結果

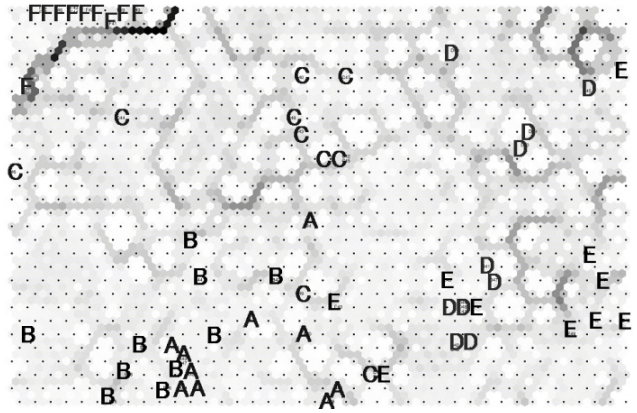

図 4　テスト用データ 60 ファイルの分類結果

7. まとめ

本研究では直感的に生成した人間乱数を解析し数列に現れる個人の特徴を探ると共に，SOM を利用した直感人間乱数データによる個人識別の可能性についての研究を行った．その結果，個人を明確に分類する指標を特定することができなかった．しかし，H，ADJ，TPI，PL，RG，RP，TKD，TIME の 8 個の指標を特徴量として同時に用い，SOM による教師なし学習を適用することにより，直感人間乱数データを概ね個人別に分類することに成功した．これにより直感的に生成した人間乱数による個人識別の可能性を追求することにした．しかしその結果は，正しく分類されたデータもあったが，明確な分類が出来なかったデータも多くあった．

これは個人内のデータにばらつきが大きいために，今回使用した指標では個人内と個人間の差を明確に分けることができなかったことが原因である．また，データの採取日にバラつきがあるなど，データ採取環境が整っていなかったことも個人内誤差の原因と考えられる．しかし，TKD，TIME の 2 つの指標については個人内誤差が少ない傾向が見られた．この 2 つの指標は他の 6 つの指標が乱数列の特徴を見るのに対し，被験者の乱数生成時の行動を見るものである．このことより直感人間乱数から個人の癖を捉えるには生成された数列の解析と同時に，生成時の被験者の行動を解析することが重要であると考えた．

以上のことより人間乱数を個人認証システムに利用するためには数列と行動 2 つの特徴を組み合わせることが有効であると考えられる．そのため，生成方法や採取環境を整備することで個人内のばらつきを減らし，またキー入力のタイミング等生成時の行動を新たな特徴とすることで個人識別の精度を高めることが今後の課題である．

8. 参考文献

[1] 乱数テスト研究会(1973), “人間乱数―頭脳のプリズム―”, 自然, 8 月号(中央公論社), pp.49-57.

[2] W.A.Wagenaar(1972), “Generation of Random Sequences by Human Subjects: A Critical Survey of Literature”, Psychological Bulletin, Vol.77, pp.65-72.

[3] 楊静宏, 川原正弘, 五味壮平, 新貝鉚蔵(2006), “人間乱数の分析”, The 20th Annual Conference of the Japanese Society for Artificial Intelligence, 1A1-2.

[4] 矢内浩文, 森太香夫(2005), “人間が生成するランダム系列の性質―シンボルの種類と生成手段への依存”, 電子情報通信学会技術研究報告, NLP2005-71.

[5] 三島雅史, 田中美栄子(2007), “短い人間乱数による診断可能性と指標の選定”, 情報処理学会論文誌：数理モデル化と応用(TOM19), pp.47-54.

[6] 榎本良太, 田中美栄子, “逆テンキー（MPK）方式による短い人間乱数”, MPS-69, IPSJ SIG Technical Report, pp.27-30, 2008.

[7] J.N.Towse, D.Nell(1998), “Analyzing Human Random Generation Behavior: A Review of Methods Used and a Computer Program for Describing Performance”, Instruments & Computers, Vol.30, pp.583-591.