Effective Indices to Characterize Short Sequences of Human Random Generations

Masashi Mishima and Mieko Tanaka-Yamawaki
Department of Information and Knowledge Engineering, Tottori University, Tottori 680-8550 Japan

Abstract

The brain impediment such as dementia is a serious problem today. It would be very useful if software for private diagnosis were available. In this paper, we show the effectiveness of the human random generation test (HRG) for such software and propose a set of four indices to be used for classifying the HRG data.

Human generated random numbers have strong characteristics compared to the computer-generated random numbers [1,2], and it is known to be correlated to the individual characters of subjects [3-6]. However, the analysis using the correlation dimension or HMM [4-6] requires a long data sequence thus not suitable for diagnoses.

We therefore focus on short sequences of HRG and search for effective indices to detect the sign of brain disability hidden in the HRG data. We study data taken from subjects of different age groups and successfully differentiated data from the different age groups.

1. Introduction

The human random generation test (HRG) is a handy way to measure the brain condition. It checks the flexibility of thinking in a simple manner without using any apparatus. Earlier it was used in the filed of clinical psychology to test the degree of advanced stage of schizophrenia [1]. Around 1970 in Japan, a number of researchers in the field of developmental psychology conducted a statistical study of HRG by collecting data from subjects of various developing stages, including the elementary school pupils, high school students and grownup adults [2]. Later we employed computer-based analysis to detect subtle difference in the data taken from normal adults [3], by using the correlation dimension as well as the technique of HMM [4-6]. However, those methods implicitly assume large-sized data thus not suitable for practical need of diagnosis. The longer the gathered number becomes, the more the subject's load increases.

Here we report our recent study on short HRG of length 50. Our purpose is to examine whether the short human random numbers can be used for early detecting of the brain impediment such as dementia, and identify effective indices to diagnose the symptom.

We examine various indices found in the literature [7] and select effective ones. We also propose a new index, RP, to classify the features of the short data sequences effectively. By using a set of four efficient indices, we have successfully classified data to the proper age groups. Assuming that the dementia is an extreme case of the shortage of memory capacity due to aging, we propose this method for diagnosis of the early symptom of brain deterioration. We further present our result in terms of self-organized map (SOM) [8,9].

2. Method of Data Acquisition

Our method of data acquisition is as follows. The subjects are asked to generate 50 numbers orally by randomly choosing one letter from {0, .., 9}. Data are orally generated by the subjects and are immediately recorded by the researchers into computers. The oral test is more suitable than the writing test for detecting the level of concentration of the subjects who tries to generate the numbers evenly by memorizing the sequence in the past. During the test, the subjects are directed to close their eyes or to see the ceiling in order to avoid external disturbance. We do not specify the speed of generations in order not to give extra pressure to the subjects.

We have taken 50 data sets from each subject of age 20s (male and female) and 10 data sets from each subject of age 30s, 40s, and 50s (male and female). In

addition, we have obtained data of three patients of schizophrenia.

3. Human random numbers

3.1 Basic Properties

We assume that the degree of complexity of the human-generated random numbers reflect the capacity of the brain. If the subjects attempt to generate every figure of 0 to 9 as well as every arrays of figures evenly, they would employ the maximum brain capacity in order to memorize as many figures as possible they have generated so far. As the subjects get tired, the concentration level would go lower Thus we use the deviation from the randomness as the symptom of the deterioration. We attempt to quantify the memory capacity by using human random numbers. We introduce the used indices as follows. In this chapter, normal person's value is average of the 20s subject's data.

3.2 Entropy (H)

The Shannon-entropy H is the first index to examine the degree of randomness. H measures the average information obtained from a sequence of letters, and it measures how random the letters are in a sequence of letters. It can be quantified as in Eq.(1) by using the probability p_i of appearance of the i-th figure.

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

H takes its maximum value when every letter appears with the same probability. Dividing this quantity by its maximum value H_{max}, we use the relative entropy for convenience. In this definition, H ranges from 0 to 1. We assume that the higher value of H means the more active memory capacity of the brain of the subject, in order to use the letters evenly.

It would be necessary to consider the evenness of the arrays of letters, because, e.g., a sequence such as "0123456789" cannot be considered as random, although the nine letters appear evenly. We need to maximize entropies of arrays of various lengths in order to measure the randomness. However, those entropies of arrays are not suitable to measures the randomness of the sequences shorter than 100, because it is too short to have all the patterns of arrays within the capacity of 100 letters. For this reason, we consider only H defined in Eq. (1)

3.3 Coupon Score (CS)

The coupon score (CS) [3] is defined as the length of sequence before all the letters (0 to 9) appear. The CS is approximately 30 on the average for machine generated random numbers. However, the average of normal person is 17, which means that human strongly wish to use every letter compared to computer programs. The CS occasionally takes very large values for those who have very strong preference in choosing specific numbers.

3.4 Turning Point Index (TPI)

The turning point index (TPI) [3] measures how frequently the switch from ascending pattern to descending pattern, and vice versa, occurs in the data sequence. The turning point (TP) is defined as the letter after which the pattern changes. For example, "135426" has two turning points, 5 and 2. The turning point index (TPI) is defined as the ratio of TP and its expected value, where m (=50) denotes a maximum data size.

$$\mathrm{TPI} = 100 \times \frac{\mathrm{TP}_{\mathrm{observed}}}{\mathrm{TP}_{\mathrm{expected}}} \qquad \mathrm{TP}_{\mathrm{expected}} = \frac{2}{3}(m-2) \qquad (2)$$

The TPI is highly vulnerable to the human brain condition. When the subjects is active, it tends to be larger than one, while for inactive subjects or patients in advanced stage of mental disease it tends to be smaller than one.

3.5 Adjacency (ADJ)

A remarkable feature of human generated random numbers is the lack of repeats of the same figures successively. Guided by this fact, we utilize the adjacency (ADJ) in order to characterize the data. Focusing on the difference between two adjacent figures, we classify the data by the absolute values of the differences (d) between adjacent figures into four types, d=0, d=1, d=2, d>2. All the data indicate that the rate of d=0 is extremely lower in human generated data compared to computer generated random numbers. Also the rate of d=1 is a good measure of mental condition. For example, the data taken from the schizophrenia patients are characterized by an excess

amount of d=1 compared to the data from normal subjects.

3.6 Null Score Quotient (NSQ)

The null score quotient (NSQ) [3] measures the degree of deviation from the even generation of pairs (array of length 2). It is defined as

$$NSQ = 100 \times \frac{NS}{a^2 - 1} \qquad (3)$$

where NS denotes the numbers of pairs not appearing in the sequence and a denotes the size of letters used. In the case of using decimal figures {0,…,9}, a=10.

This is a useful index for long sequences of HRG, since the differences between subjects/conditions in HRG reflect nicely in this index. However, NSQ is not suitable for the data of length 50, which is too short to contain all the patterns of pairs for a=10.

3.7 Repeat Pattern (RP)

We propose a new index to be used in place of NSG for the case of short HRG. Since the subjects of HRG try to generate the next letter based on their memory of the last generated letter, NSQ is a good measure for the memory capacity of the subjects. However, the problem is that the value of NSQ ranges from 51.5 to 100 for the case of data sequence of length 50. We need a better index for short data.

Consider the case when the generated data is “1358763” so far, and 5 is about to come out next, one would make an effort to avoid 5, by considering the previously generated 35. Human would pay all the effort to improve the randomness (complexity, in fact). Guided by this thought, we define a new index as the frequency of repeated pairs, by Eq.(4).

$$RP = 1 - \frac{NRS}{m - (n-1)} \qquad (4)$$

where NRS denotes the number of unrepeated pairs, m denotes the length of the sequence(m=50), and n denotes the length of array(n=2 for pair). The more the repeated pairs, the larger the value of RP, indicate the deterioration of the memory capacity of the subject. We have also studied the case of n=3. The result was not very different from the case of n=2(pair). Thus we stick to the case n=2 and consider only the case of unrepeated pairs.

Note that RP ranges from 0 to 100, irrelevant to the size of the data sequence, unlike NSQ.

4. Study on the different age groups

We apply those indices introduced in the previous chapter on various data including schizophrenic patients, normal subjects in different age groups(20s, 20s, 40s, 50s and over), and computer-generated random numbers.

Figure 1 shows the mean values of indices, RP, NSQ, TPI, ADJ, CS, H. for 20 subjects, in the order of ascending age groups. The subject number from 1 to 10 belong to the age group of 20s, the number from 11 to 13 belongs to the age group of 30s, and the number from 14 and 15 belongs to the age group of 40s, and the number from 16 to 20 belongs to the age group of 50s and over. We do not distinguish the actual ages within an age group.

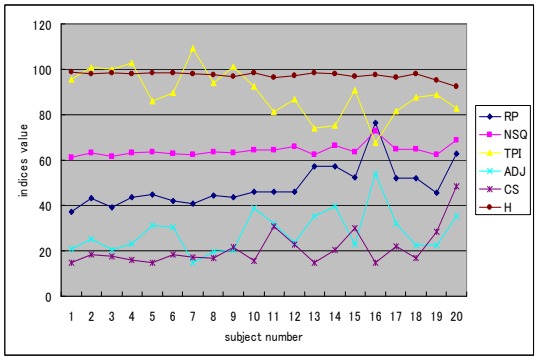

Fig. 1. Values indices vs. individual subjects

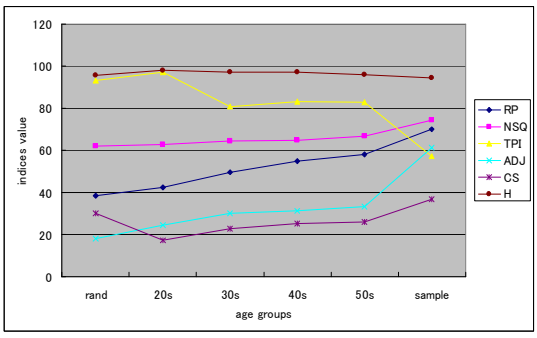

Fig. 2. The value of indices vs. age groups

Fig.2 shows the mean values of the same indices as in Fig.1 calculated for each different age group. In both figures, the entropy H show the relative value to its maximum, and are expressed in terms of percentage.

In both figures, the values of indices show strong correlation to the age groups. This fact suggests the possibility of using HRG for dementia diagnosis. It also indicates the effectiveness of the indices used in our work.

5. SOM for multi-dimensional data

The indices used so far are mutually dependent. We compute the Pearson product-moment correlation coefficient between indices and show the result in Table.1.

Table1. correlation coefficients

	RP	NSQ	TPI	ADJ	CS	H
RP		0.97	-0.25	0.34	0.13	-0.3
NSQ			-0.29	0.39	0.15	-0.33
TPI				-0.6	-0.07	0.1
ADJ					0.02	0.04
CS						-0.66

According to the result shown in Table.1, there is a strong positive correlation between NSQ and RP and the information from those two indices are overlapping. We prefer RP to NSQ because the range is larger in RP for short data of length 50 thus more suitable to observe the difference.

A negative correlation exists between H and CS. We prefer H to CS based on the following observation. The value of CS can easily become large if the subject forgets to say one letter no matter how random the other parts of the data are. It is rather hard to distinguish between a completely regular sequence but missing one letter, and a highly complex sequence but missing only one letter.

Thus, we are left with four indices, RP, TPI, ADJ and H. However, they are still mutually dependent. One good way of presenting multi-dimensional data is given by the self-organizing map (SOM), which works well for the case of mutually dependent multiple variables. The result is shown in Fig.3, where A, B, C denote the age group of 20s, 30s+40s, and 50s+up, respectively. All the indices are normalized to be in the range of [0, 1]. The parameters of SOM are chosen as follows. The Map size is 15×10, learning coefficient α is 0.5, the initial radius is 9, and 10000 cycles of learning.

Fig. 3 shows the data are properly classified by means pf those four indices. A denoting the young group of age 20s aggregate to the left, while C denoting the older group of age 50 and higher aggregate to the right, and B denoting the middle age group are located in between.

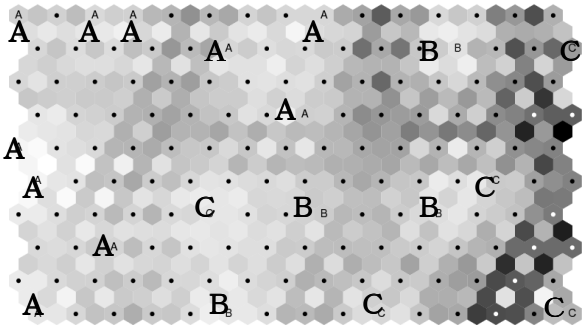

Fig. 3. RP, TPI, ADJ and H for data from three age groups, A(20s),B(30s,40s),C(50+) represented by SOM

6. Conclusion

We have examined indices applicable for short HRG in order to diagnose dementia, and selected a set of four effective indices including RP, which we have proposed in this paper. We have also presented the result by SOM. We have successfully classified the data of different age groups, conditions, etc. according to their characteristics by using the indices of our choice.

References

[1] Early works are critically reviewed in: Wagenaar WA, (1972), Psychological Bulletin 77, pp.65-72.

[2] Ransuu Test Kenkyuukai (Society of Human Random Generation Tests) (1973) Shizen 1973-8. pp.49-57

[3] Iba Y, Tanaka-Yamawaki M (1996), Proc. 4th Int. Conf. Soft Computing IIZUKA'96, pp.467-472.

[4] Tanaka-Yamawaki M, Masuda H, Kawagoe M (1998), Proc.3rd AROB, pp. 610-613

[5] Tanaka-Yamawaki M (l998), Proc. ICONIP'98, pp.215-218;

[6] Tanaka-Yamawaki M (1999), Human Generated Random Numbers and a Model of the Human Brain Functions; Proc.1999 IEEE SMC, pp. 223-228

[7] Towse JN, Nell D (1998) Analyzing human random generation behavior, Behavior Research Methods, Instruments, & Computers,30(A), pp.583-59

[8] Kohonen T (1995), Self-Organization maps (in Japanese)

[9]SOM_PACK http://www.cis.hut.fi/research/som-research/