Application of SOM in a health evaluation system

H. Tokutaka[1], Y. Maniwa[2], P.K. Kihato[3], K. Fujimura[3] and M. Ohkita[3]
1) SOM Japan Co-LTD, 2) Futaba Clinic, Osaka, 3) Electrical and Electronic Department, Tottori University
Email: {tokutaka@somj.com, maniwayo@gold.ocn.ne.jp, kamitazv@yahoo.co.uk, fujimura@, &
mohkita@ele.tottori-u.ac.jp}

Keywords: Self-Organizing Maps, Visualization, Health evaluation

Abstract**—** A health evaluation system was constructed which visualizes the living habits and health state from a person's checkup list by using the feature of SOM that multi-dimensional data can be mapped onto a two-dimensional surface. Here, three examples cases are reported. A change to the health region of the map by taking medication was visualized by the SOM from the conventional numerical expression. Also, the specific sick record converges towards the sick region of the map when the disease progresses. However, it was shown and visualized from the sick records that no convergence occurred in the case of metastasis of cancer even if for the same examinee, the cancer had progressed. Finally, for the display of the health point mark, and the display of the sick record, the spherical surface SOM, is demonstrated to be suited in the visualization.

1 The Research Background

At present, for the Japanese, there are few subjective symptoms in lifestyle diseases such as serious high blood pressure, arteriosclerosis, diabetes and so on. The early detection of the disease and the early treatment were thought as the purpose of the physician's examination. In recent years, the nature of the physician's examination is changing, as it becomes the starting point to investigate one's health state by reconsidering lifestyles such as eating habits and physical exercise habits. That is, the physician's examination of any lifestyle disease (too much alcohol consumption can cause liver damage) is ordinarily and generally performed and within the monitored parameters, a particular parameter is almost established [1]. The purpose of the physical examination is to have an early forecast and project improvement of the lifestyle disease. In the conventional physical examination, only the result whether the value of the inspected parameter was normal or not was reported, or to what extent it was extraordinary. For the general public with little medical knowledge, it was almost unclear to what clinical condition or what kinds of living habits need improvement. In this research, the multi-dimensional data can be represented in the two dimensional plane using the characteristic of the SOM [2]. Using the SOM, the relation between the living habit and the health can be visualized from the health certificate of the individual. Thus, the relation between each inspection and the clinical condition can be made clear. Thus, the health evaluation system, which can display the health condition, can be constructed. An outline on how health SOM map is constructed has already been reported in WSOM03 [3].

2 The health evaluating SOM

2.1 About the physical examination data

The SOM was trained on the medical checkup data of 394 examinees with the cooperation of X-city medical center and Y doctor. The following parameters were used for the analysis: BMI (Body Mass Index) which were measured from weight and length, the systolic blood pressure (blood pressure high), the diastolic blood pressure (blood pressure low), the total cholesterol, LDL cholesterol, HDL cholesterol, triglyceride (TG), GOT, GPT, ChE (the cholinesterase), HbA$_{1C}$ (the hemoglobin A$_{1C}$), γ-GTP, the uric acid, the urea nitrogen, the creatinine, the hemoglobin Ht (the hematocrit). Thus, there were 17 parameters in total. Each inspected parameter has a normal value or range. Also, GOT, GPT are the deviation enzyme of the liver and are the index of the hepatopathy. A certain medical relation for every parameter is decided. Referring to Tables 1 and 2 in [3], normal values for every parameter and their medical relation can be realized respectively. Using these data, a usual plane SOM was trained. Also, data from 4000 female examinees (specifically, no alcohol addiction), 3000 male (specifically, no alcohol addiction), and 700 male (with drinking habits) all from 1999 to 2003 were used. The following 11 inspection parameters were used, which were GLU (the blood sugar level), HbA1c, AST (GOT), ALT (GPT), TG, LCHO, BMI, UA, H-BL (the systolic blood pressure), L-BL (the diastolic blood pressure), and γ-GTP. GLU was added as the risk factor to the diabetes in addition to Table 2 in [3] with its normal value being between 70 and 109. Also, when either or both of AST, ALT, exceeded a normal value, AST>0.87ALT or AST<0.87ALT was distinguished.

X, Y* and Z* Names withheld.*

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

2.2 Data pre-processing

When creating the SOM map, one must deal with the fact that the units of measurement differ from parameter to parameter. Hence, the following normalization procedure was applied. Let us take a minimum normal value as L, a maximum normal value as H, the data value as X, and the normalized value as Y:

when $(X < L)$;

$$Y = X / L \tag{1}$$

$(L \le X \le H)$;

$$Y = 1 \tag{2}$$

when $(X > H)$;

$$Y = X / H \tag{3}$$

Equation (2) refers to normal health range whereas Equations (1) and (3) give linear scaled values of out of normal range values. However, after pre-processing the data, some parameter has sometimes an extremely high value. Since it is our purpose to evaluate a general lifestyle disease (too much alcohol can cause liver disorder), such a high value is not desirable; we decided to use a ceiling-value for these types of normalized data values.

Frequency distribution for every parameter was produced to obtain a ceiling-value of equation (3). As shown in [3], most people were distributed within 1.5 of the normalized value in any parameter except the triglyceride (TG) and γ-GTP. The people with the normalized value more than 1.5 ceiling happened to still be within normal distribution in case of the triglyceride (TG) and γ-GTP [1, 3]. A higher ceiling value of 2.0 was therefore used in case of the triglyceride (TG) and γ-GTP with the other parameters' ceiling-value fixed at 1.5 as was originally selected. Any normalized value exceeding ceiling values was given the ceiling value.

2.3 The construction of the SOM map

Using preprocessed data, SOM map of the physical examination data can be constructed where several learning conditions can be examined. Prior to the construction of the SOM map, the conditions that most suited for the visualization of the physical examination data were searched for and the SOM map constructed. In case of the data of the 394 examinees with the first 17 parameters, the normal health plane SOM map was used.

Also, for the construction of the map, weights were not used in the case of the 17 parameters of Table 1 [3]. However, in the case of the following 11 parameters, the following weights were imposed on the important parameters. The weight was 2 for GLU (the blood sugar level), HbA1c, AST (GOT), and ALT (GPT). The weight was 1.5 for TG, and LCHO. For all the others, 1.0 was used. The weight of each element was based on its importance to the health state. Here, Torus type SOM was used for the map construction [4].

In health evaluation, which is the purpose of this article, every clinical condition must be classified, and it is also necessary to make the boundaries clearly. Therefore, since each inspection parameter has a medical relation as shown in Table 2 [3], a class area of every clinical condition can be fixed using this relation. That is, the abnormal place of the component map of every parameter, which is related with the clinical condition, is selected and the area is fixed. For example, let us consider the relation between the component map of the parameter related with the kidney disease and the health evaluation SOM map. It is generally confirmed that the positions in the SOM map of high urea nitrogen and high creatinine in the component maps, are related to the kidney disease. The other medical relations using SOM and the other component maps can be examined [3]. In this way, the health evaluation SOM that can display all medical relations can be constructed. The SOM map emerging from these considerations and constructed using the former 394 examinee's data is shown in Fig. 1.

2.4 The construction of the Torus type SOM map

The three types of SOM maps, Plane, Spherical and Torus have their merits. Torus type is used mainly due to co-relation of data at the edges and corners and at the same time viewing the entire lattice. Thus the whole spectrum of data can be looked as health class and the monitored parameters as health destructors.

The result of the Torus type SOM [4] that was constructed using input data with 11 parameters is reported. The choice was focused on retaining the most important health parameters. As for the tool, this time three sheets of the Torus type SOM maps were prepared using the data from the male (specifically, no alcohol addiction), the other male (drinking habits) and the female patients. The normal values that are related to Table 1 in [3] could separately be prepared for the male and also the female patients. The Torus SOM map for the male patients shown in Fig. 3 (a) and (b) looks different; (b) is obtained by shifting (a) to the right and slightly shifting (a) to the above. Then, it can be understood that the shifted part appears continuously on the right and also on the left. The upper and the lower part, also appear to tally in the same way. Fig. 3 is more detailed SOM compared to Fig. 1. Complicated symptoms of various diseases can be analyzed using this SOM. Here, 3000 male examinee data were used. Related cluster diseases can be visualized in detail.

3 Results and Discussions

3.1 Health points

This time, the evaluation of the health mark point is expressed by equation (4). Incidentally, in the later described example and Fig. 2, a point mark is displayed in 10-point units. Therefore, the point from 95 to 100 are rounded off and displayed as 100. It is worth noting that if an examinee had all the observed parameters being worst values, then health points would be 0. If the examinee had normal values, then health points would be 100. Any other variations would be between the two limits.

$$Health\ mark\ \ po\mathrm{int}_i = \frac{\sqrt{\sum_{n=1}^{n}(WV_n - NV)^2} - \sqrt{\sum_{n=1}^{n}(\chi_{ni} - NV)^2}}{\sqrt{\sum_{n=1}^{n}(WV_n - NV)^2}} \times 100 \qquad (4)$$

Where: WV_n is the worst value of respective parameter, NV the normal value, χ_{ni} the data of the examinee and 'n' the number of parameters (17) as in table 1 [3].

The mark in the upper right corner (the critical-region) was low in all 600 positions in the SOM of Fig. 1. Referring to the same figure, let us make the upper-left corner as the first. The start of the 2nd line from the upper-left corner becomes 31st. The upper-right corner (the 30th) is 30 mark points and then, the place (the 29th); one ahead was 30 mark points. The 2nd line from the upper-right corner (the 60th) shows the minimum point marks of 20. Moreover, the place of 1 line below (the 90th) also shows the same lowest point marks as 20. By the way, the health area where all normal values show 1 is the place shown as 100 (node number 463). The other corresponding numbers are 492, 493, and 494.

3.2 Application of plane health SOM (example) as in Fig. 1

The tool of the health evaluation map, which was constructed by using the plane SOM, can be described as follows: The value of the following year from the current one can be estimated by inserting the checkup data. For example, there are 3 years data of $H(t_1), H(t_2), and\ H(t_3)$ then using linear extrapolation; the following year $H(t_4)$ can be estimated by equation (5). Intervals t_1, $t_{2,}$ and $t_{3,}$ are known and t_4 is t_3 + 1. The estimation is agreeable for following year only.

$$H(t_4) = H(t_3) + \frac{H(t_2) - H(t_1)}{t_2 - t_1} + \frac{H(t_3) - H(t_2)}{t_3 - t_2} \qquad (5)$$

The results are displayed and the state of the movement of the health pattern of the examinee can easily be understood on inspection (see the results of Fig. 1). If patient's health state is to improve from a worse value to the normal value state, this position movement in the health area on the SOM map can be visualized (see the trend from 99 to 04 in Fig.1). Usually the doctor describes such a trend to the examinee by, saying, "Do exercise well and lower your high TG value". If the doctor shows the health trend as a movement on the health SOM map simultaneously with his explanation, it seems to be easy for the examinee to understand what needs to be done.

All of the medical examinations data of the examinee can be displayed on the SOM map, and the positions linked by the shortest distance like TSP (Traveling Salesman Problem) method (see Fig. 3). Besides, it is possible to "save 50 SOM maps at once and view them in series if need be" from the 50 examinees data of the medical examinations of the day, when the data are saved in the csv form. There are other functions to which explanations are however, omitted. Here, it is shown by some examples that the SOM is valid.

The consent from Mr. X is obtained. Let us explain how the visualization of the result will be easy for him to understand his physical checkup. Mr. X had a high acylglycerol as it turned out from the medical examination of his family doctor. Referring Fig. 2 the checkup data in 1999 comes out much higher than the normal values in parameters TG and LDL cholesterol. Fig. 1 shows this region as "obesity region" and his health point mark is low with 60. Mr. X was advised by the doctor to "take medication because the acylglycerol was very high". Two (2) years under medication, his health state in 2001 is shown in Fig. 1. His TG value improved and his position in the map moved to the area 01 of "high heperlipidemia ". The health point mark also increased to 80 (see Fig. 2). Moreover, Mr. X continued to take medicine, and his health state as of year 2002 was as shown in Figs. 1 and 2. TG value fell to the normal value and his position in the map moved to normal health area with 100 mark points (see Fig. 2).

The diagnosis of the doctor is "The value falls. You don't need to be worried but continue with the medication". The doctor admits, "Obesity must be cancelled by exercises. If the value falls, it is OK, but in this case, the fall is due to medicine. Then, continue to take medicine for maintaining the low value". The only information available is the blood data. Acylglycerol value was specifically shown to return to normal value. Mr. X was relieved. The author got the checkup data from Mr. X and showed his values of 99, 01, 02, and 04 on the map as shown in Fig. 1. Mr. X was very excited, saying “bravo". He said, "When the data is shown by such visualization, it was very easy to understand the results".

The result of Mr. X in 2004 is with the LDL cholesterol and the TG value falling as shown in Fig. 2. Fig. 1 shows this trend from position 04 to health area with 100-point mark. He was less worried since only, the value of CHE in year 04 showed to be a bit rising. In his opinion, "Originally, the liver was not robust and the fatty liver was observed by echography. His interpretation was that the bad value in his physical condition 2004 was coming by chance". Even though some of the displayed figures are displayed in red they mean to exceed normal values, 100-point marks are shown in the map representing normal health regions in the SOM map

3.3 Application of the Torus type SOM as the Fig. 3

The result of the Torus type SOM [4], which was constructed using input data with 11 parameters with important parameters weighted, is reported next. In addition to these 11 parameters, others like urea nitrogen (UN) and creatinine (CRE) were used for checking the kidney problem. Hemoglobin (Hb), and the hematocrit (Ht) were used for the identification of the anemia. All these 4 parameters were not used for the construction of the map but used individually for the identification. After the construction of the map, about 30 of the recorded sick persons’ data were extracted randomly from about 5000 data of a certain company in the authority of doctor Z, an industrial-physician, and these pieces were checked using the constructed SOM map. The types of diseases were namely: cancer, cerebrovascular accident, and heart disease. Two figures in the case of the heart disease are shown below. As for Fig. 3(a), the health mark is high with 81 points (Torus type SOM, equation (4) is used for the calculation of the health point mark with no rounding). In the case of Fig. 3(a), the looseness index is as large as 83.7.

The looseness index in Fig. 3(a) is defined as follows: The 5 points of the measured 5 years are linked at the shortest distance and the total distance is computed as the summation of the square of the distance from point to point for simplicity. A unit distance was taken as one division of a square in Fig. 3. The distance is averaged over the five (5) years (Note that the result in 97 and in 01 in Fig. 3(a) is in the same position). Fig. 3(b), the health mark is low with 50 points for the same heart disease of Fig. 3(a). The looseness index is small as 1.6. These results were arranged for every sick person example. The average combination year distance is calculated by linking all of the measurement years that were used. Also, an average health mark from all the used measurement years is used for the index of the health point mark.

All data are arranged and shown in Fig. 4. The results are discussed as follows: A diagnosis is being cleared like the heart disease, the cerebrovascular accident, and so on. Then, when the health mark is low, there are "convergences" and few movements in the specific area of the independent disease. That is, "the looseness index" decreases. However, in the case of cancer, there is not a specific diagnosis. For example, the location moves to another area on the map if another area of the body is attacked via "the lymph node". Also, the location on the map doesn't move of a person whose area of cancer is fixed on a specific part of the body. Such cases pile up and there are no correlations between "the average health mark" and "looseness index". Therefore, for the person whose "average health mark" is low and the "looseness index" is large, the suspicion of cancer will remain. Also, it is possible to be used for the estimation of the diagnosis of the examinee, too, oppositely if the area, onto which each sick person example converges, can be specified. In the case of the individual sick person’s record, the diagnosis can be estimated from the convergence of the position on the map. For the person whose "average health mark" is low and "looseness index" is large, the follow-up care will be necessary by supposing that there is suspicion of cancer. Incidentally, for the healthy person whose "average health mark" is always 100 points, his location on the map will not be loose and not far from the health position. Therefore, the result should converge like the dotted line on Fig. 4 (b), (c), and (d).

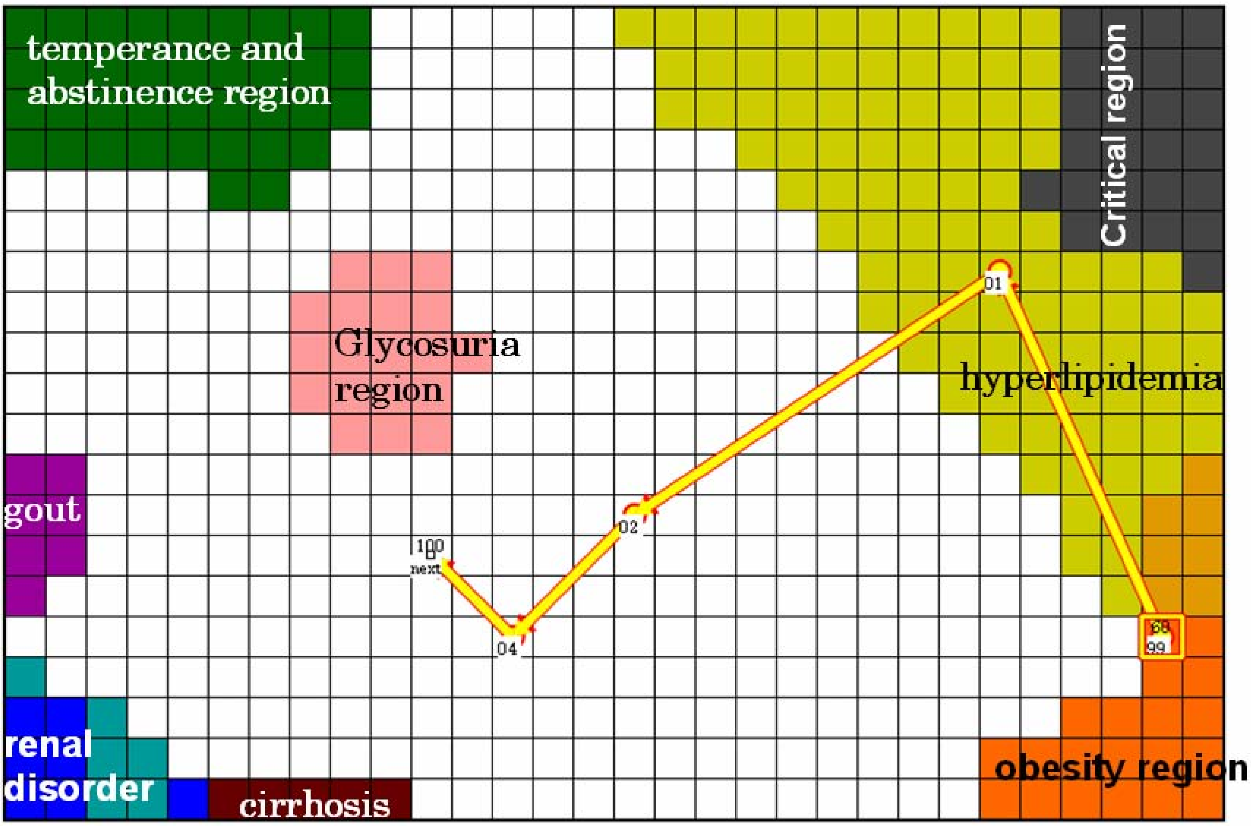
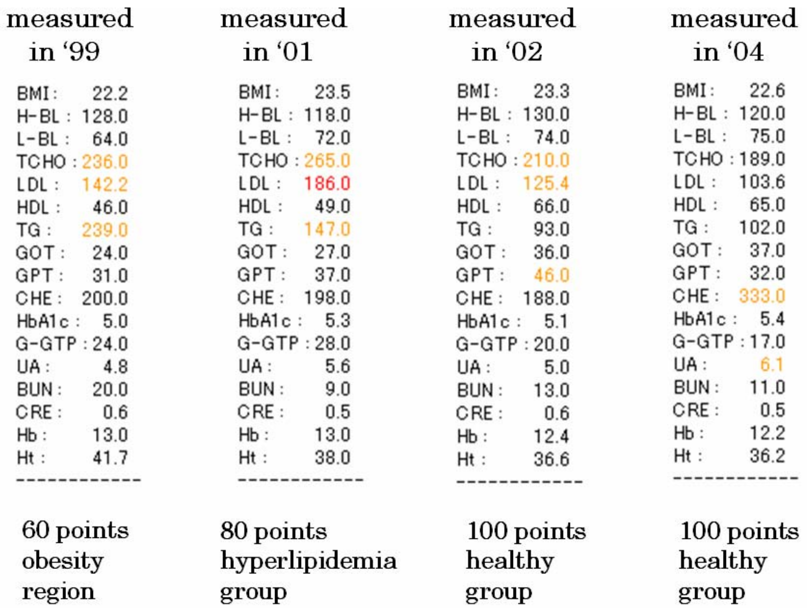


Fig. 1 The health-check SOM map by the 394-examinee data. 01 is the position of 2001 of Mr. X. who began to take medicine for 2 years. In 2002, it then moves to the position of the health area as 02. 2 years later, the position was moved to 04 of health area. Here, "next" is the estimate position in 2005 using the data for these 4 years and the equation (5).



measured in '99		measured in '01		measured in '02		measured in '04	
BMI :	22.2	BMI :	23.5	BMI :	23.3	BMI :	22.6
H-BL :	128.0	H-BL :	118.0	H-BL :	130.0	H-BL :	120.0
L-BL :	64.0	L-BL :	72.0	L-BL :	74.0	L-BL :	75.0
TCHO :	236.0	TCHO :	265.0	TCHO :	210.0	TCHO :	189.0
LDL :	142.2	LDL :	186.0	LDL :	125.4	LDL :	103.6
HDL :	46.0	HDL :	49.0	HDL :	66.0	HDL :	65.0
TG :	239.0	TG :	147.0	TG :	93.0	TG :	102.0
GOT :	24.0	GOT :	27.0	GOT :	36.0	GOT :	37.0
GPT :	31.0	GPT :	37.0	GPT :	46.0	GPT :	32.0
CHE :	200.0	CHE :	198.0	CHE :	188.0	CHE :	333.0
HbA1c :	5.0	HbA1c :	5.3	HbA1c :	5.1	HbA1c :	5.4
G-GTP :	24.0	G-GTP :	28.0	G-GTP :	20.0	G-GTP :	17.0
UA :	4.8	UA :	5.6	UA :	5.0	UA :	6.1
BUN :	20.0	BUN :	9.0	BUN :	13.0	BUN :	11.0
CRE :	0.6	CRE :	0.5	CRE :	0.6	CRE :	0.5
Hb :	13.0	Hb :	13.0	Hb :	12.4	Hb :	12.2
Ht :	41.7	Ht :	38.0	Ht :	36.6	Ht :	36.2
------------		------------		------------		------------	
60 points obesity region		80 points hyperlipidemia group		100 points healthy group		100 points healthy group	

Fig. 2 the checkup data of Mr. X in 4 years that were inputted to the SOM map. The decrease of TG was confirmed by taking medication.

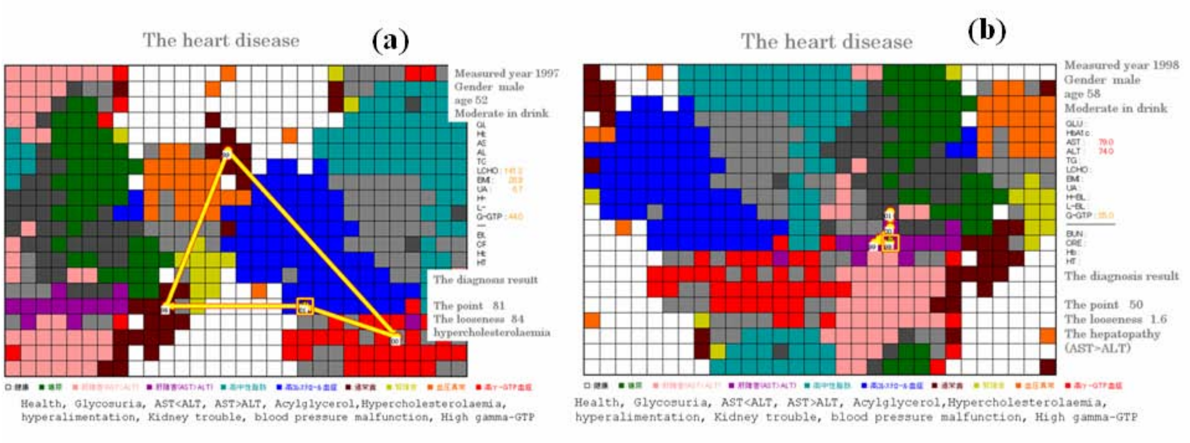


Fig. 3 Torus type of SOM map

Generally, the Torus type map can be moved freely in the top and the bottom, and also on either side. Map (b) was obtained by moving the data of the examinee of map (b) to come to the center of map (a). The shortest distance links the 5-year record from 97 to 01. Incidentally, the normal value of the each parameter is painted in white. The health mark is high with 81 and the looseness index is also high with 84 for map (a). In map (b), the health mark is low with 50 for another examinee of the heart disease, with a looseness index as small as 1.6.

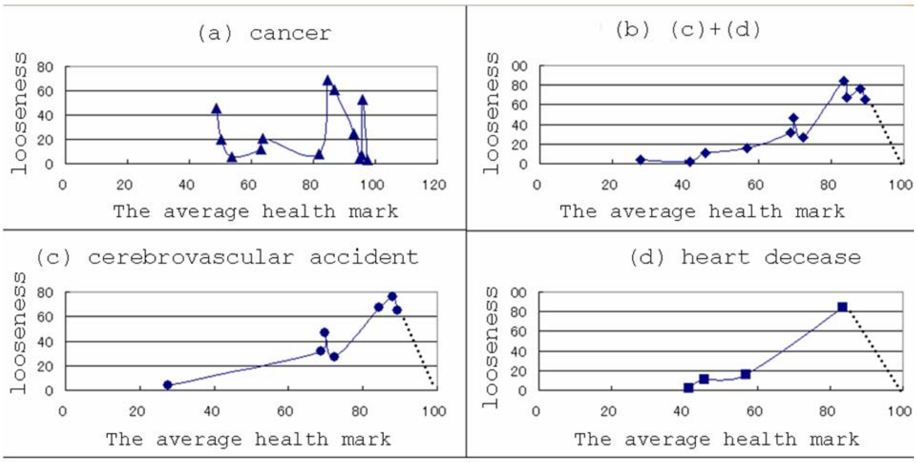


Fig. 4 an average health mark and looseness index was arranged in each example.

As for the cancer, no correlation is observed. However, for other 2 diseases, the looseness index becomes small, when the mark becomes low. Incidentally, the dotted line in the figure is the estimate of the healthy person.

3.4 Construction of the spherical health SOM map

Next, the spherical surface SOM [5,6] where the phase expression is expected as uniform is applied to this health SOM map. Using the 2400 data of males with no alcohol addiction and the spherical surface SOM tool "blossom" [7], the health map is displayed on the spherical surface. The results are shown in Figs. 5-7. Figures 5 and 6 are U-matrix representation of spherical SOM where white color is healthiest zone and dark gray zones representing most risky zones. In Fig. 5(a), a health area is displayed at the center with no labels but the healthiest area indicated with node “HH" with its neighborhood health areas labeled “H". Fig. 5(b) is the opposite surface of Fig. 5(a). The area with minimum health-point is shown with the sick label. Labels in Figs. 5-7 are as follows: G: diabetes, L: liver disease, T: TG arteriosclerosis, C: LDL cholesterol, γ: high γ-GTP. Fig. 7 shows component maps. Here blue is healthy zone and red risky zone.

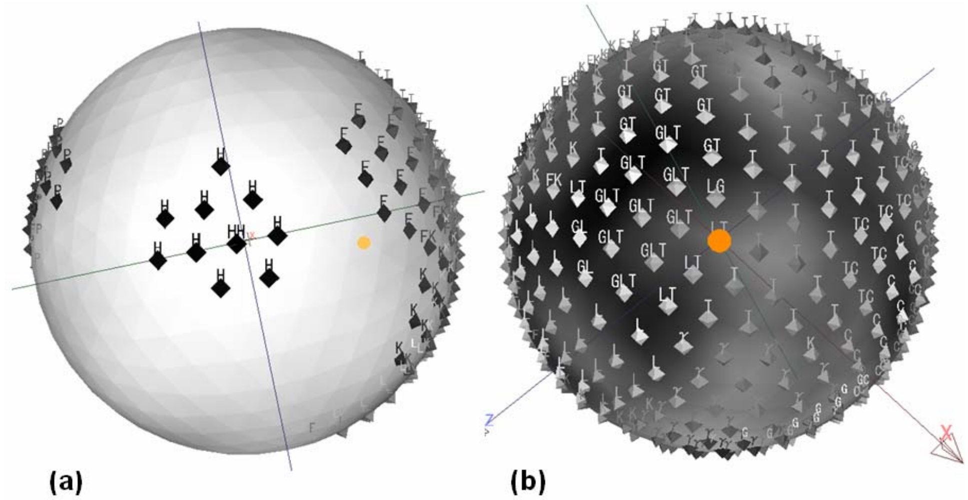


Fig. 5(a) healthiest place labeled "HH". Fig. 5(b) is the reverse of (a) with center orange marked.

4 Conclusions

In this research, physical examination data were analyzed as an application to the medical field of the SOM. The medical relation which was empirically obtained could be indicated visually by constructing a health evaluation SOM map. The health evaluation by the map could be proved to be possible. It is thought that a better precision will be obtained by analyzing a wide range of medical checkup data. In adding a feature like plane SOM, the application to a wide range of fields can be expected for a spherical surface SOM in the future.

Finally, it can be understood that by representing a plane SOM into a spherical surface SOM, the phase of each position can be represented correctly. By the fact, the health degree of each position of the plane SOM map can be displayed quantitatively by the point mark using equation (4). Thus, it will be thought that the more quantitative health evaluation by the analysis of the health evaluation data using SOM explained by this article can contribute to the further development of the medical field.

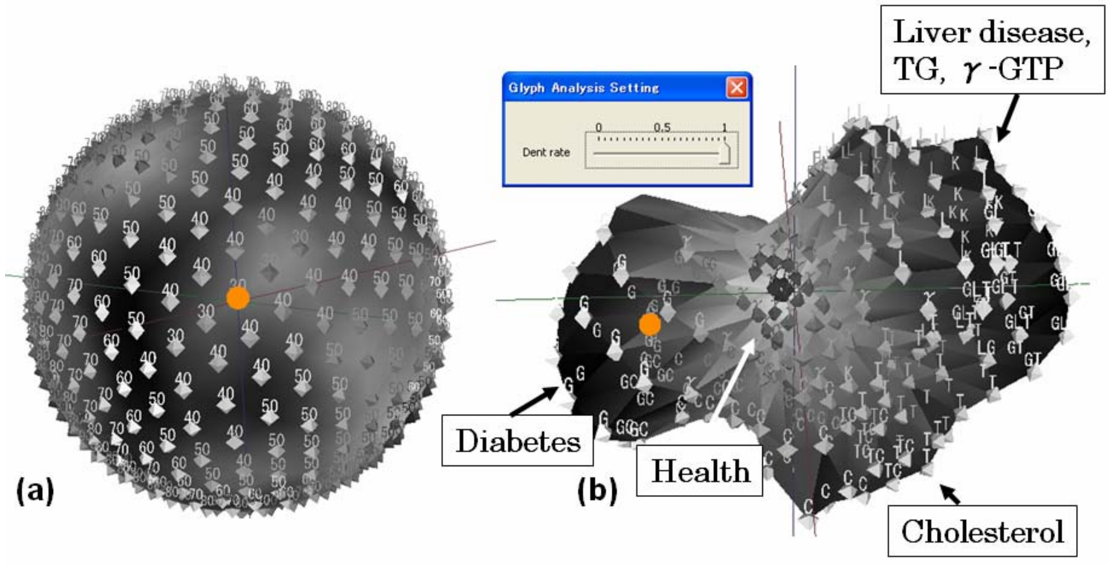
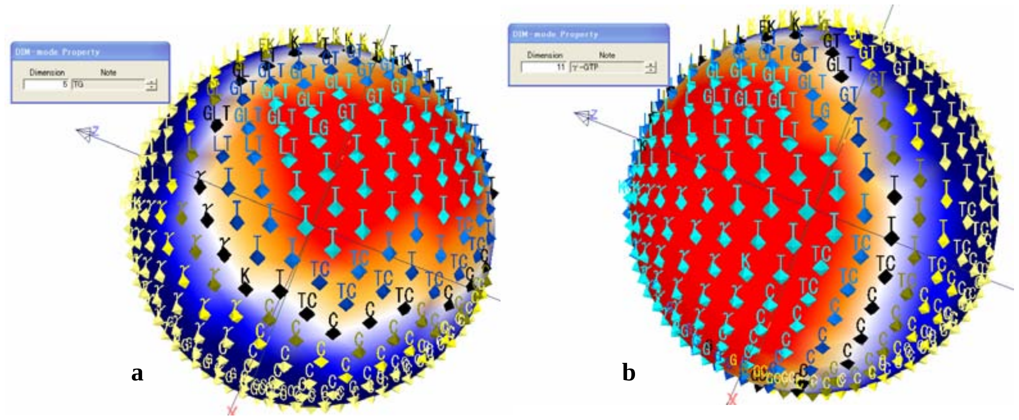


Fig. 6(a) The lowest point area on the spherical surface of Fig. 5(b) is displayed by setting “Glyph Analysis Setting” "blossom"[7] at 1. "Health" is in the neighborhood of the spherical center at its lowest. Diabetes is on the left. Liver disease, TG and γ-GTP are respectively large and on the upper right. Cholesterol is drawn as the small "bump" at the bottom.



DIM-mode Property

Dimension: 5 Note: TG

DIM-mode Property

Dimension: 11 Note: γ-GTP

Fig. 7 Two-component maps display the lowest point area of Fig. 5(b) and Fig. 6(a) are high. Components are (a) TG and (b) γ-GTP.

References

[1] I. Kanai, original, M. Kanai, editing: “The necessary proposals to the clinical examination method -the 31th revised edition-”, in Japanese, Kinbara publishing Co. Ltd, 1998.

[2] T. Kohonen, “ Self-Organizing Maps”, *Springer Series in Information Sciences,* Volume 30, 2001.

[3] H.Kurosawa, Y.Maniwa, K.Fujimura, H.Tokutaka, and M.Ohkita, “Construction of checkup system by Self-Organizing Maps”, *Proceedings of Workshop on Self-Organizing Maps* (WSOM'03), pp.144-149, 2003.

[4] H. Tokutaka, M. Ohkita, and K. Fujimura, editors, “The Self-Organizing Maps and the Development--From medicine and biology to the sociological field. (a tentative title) Appendix (B)”, in Japanese, Springer-Japan Inc. to be published in April 2007.

[5] H. Ritter, “Self-Organizing Maps on non-Euclidean Spaces”, *Kohonen Maps*, Editors, E. Oja, and S. Kaski, Elsevier, pp.95-110, 1999.

[6] D.Nakatsuka and M. Oyabu, “Application of Spherical SOM in Clustering”, *Proceedings of* Workshop on Self-Organizing Maps, (WSOM'03), pp.203-207, 2003.

[7] http://www.somj.com/