Fourth International Workshop on Computational Intelligence & Applications
IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, December 10 & 11, 2008

User's Comment Classifying Method Using Self Organizing Feature Map on Healthcare System for Diabetic

Kazuya Mera and Takumi Ichimura
Graduate School of Information Sciences, Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, Japan
email:mera@hiroshima-cu.ac.jp

Abstract— **Diabetes is a metabolic disorder characterized by the elevation of blood glucose. Glysemic control can delay the onset and slow the progression of vascular complications. Lifestyle modification including weight reduction can contribute significantly to glycemic control. The Health Support Intelligent System for Diabetic Patients (HSISD) can provide guideline-based decision support for lifestyle modifications in the treatment of diabetes. HSISD also provides opportunities for telecounseling (TC) with the use of mobile devices and the Internet. The telecounseling phase inquires about the patient's condition and the patient answer in a questionnaire. In the questionnaire, there is a question like "Have you developed any symptoms of anxiety? If yes, tell me the details." The answer is described freely so the physician should read all of patient's answer. But it is hard for physicians to read all text carefully because a physician has a lot of patients. We propose a method to analyze text data from the patients and classify them into five anxiety types (mental problem, physical problem, diet, physical activity, and medicine) automatically. Related to the classified anxiety type, the method can analyze the patient's inner emotion to guess serious and emergency degree of the patient. In this method, Self organizing feature map is trained by the distribution of feature words (morphemes) in the input text and also classifies anxiety type and emotion type.**

I. INTRODUCTIONS

The incidence of diabetes is increasing worldwide. Diabetes is a metabolic disorder characterized by the elevation of blood glucose. It increases the risk of vascular complications, which are the most important causes of morbidity and mortality related to the disease. Glycemic control (lowering blood glucose level to the normal range) can delay the onset and slow the progression of vascular complications [1]. Lifestyle modifications, including weight reduction, can contribute significantly to glycemic control. Diabetic patients need to make a lifelong commitment to maintaining a healthy lifestyle.

Attending physicians (usually general practitioners) are at the forefront of prescribing lifestyle modifications and drugs for individual patients. Physicians assess data collected from various sources, identify problems, set specific goals, and make patient-specific plans, on the basis of knowledge gained from clinical practice guidelines (CPG) and the physician's own experience and beliefs. To enhance patient's adherence to the plan, they need to take into account not only guideline-based recommendations but also the needs and abilities of individual patients.

Planning behavior change (lifestyle modifications) can be divided into the following three steps: (1) identifying the current health behavior problems, (2) setting specific treatment goals, and (3) assigning priorities among the goals. Each goal should be derived from guideline-based recommendations when possible and expressed in a concrete and patient-specific value [2]. Ideally, a final plan should consist of two or three high-priority goals. Patient's readiness to change should be taken into account when assigning priorities among the goals [2, 3].

We have already proposed the Health Support Intelligent System for Diabetic Patients (HSISD [4]), which can provide guideline-based decision support for lifestyle modifications in the treatment of diabetes. HSISD also provides opportunities for telecounseling with the use of mobile devices and the Internet.

HSISD has two functions: GDS (Guideline-based Decision Support) and TC (TeleCounseling). The GDS function identifies areas for improvement and generates patient-specific recommendations for lifestyle modification. The TC function inquires about the patient's actual condition and offers advice to the patient by interactive communication devices. Most of the questions are filled by selecting the values and checking the boxes. The answers are utilized to calculate the patient's condition automatically. On the other hand, the patient can declare his/her detailed anxiety in the text description during TC phase. The physicians have to read all text from patients carefully because to extract explicit patient's intention from free-style description is difficult for computer system. Then it will support that the physician understands the text to analyze the type of anxiety topic of the patient in advance. Furthermore, it supports guessing serious and emergency degree of the content of the description to analyze the mental condition of the patient when he/she describes the text.

In this paper, we develop the system to be able to analyze free-style description about the patient's symptom or anxiety by using Self organizing feature map (SOM [5]). We also analyze the text data in the healthcare network community for diabetic patients as seen in SNS (Social Networking Service [6]). The method analyzes the topic type of patient's anxiety (mental problem, physical problem, diet, physical activity, and medicine) and the inner emotion of the patient (happy, sad, fear, and anger). The HSISD system proposes the classification result with an adaptive healthcare plan to the patient.

II. HEALTH SUPPORT INTELLIGENT SYSTEM FOR DIABETIC

A. Health Support Intelligent System for Diabetic

Health Support Intelligent System for Diabetic Patients (HSISD) [4] focuses on lifestyle modifications in type 2 diabetic patients. The predominant users of this system are general practitioners who care for type 2 diabetic patients. As shown in Figure 1, HSISD has two functions: GDS (Guideline-based Decision Support) and TC (TeleCounseling). For the GDS function, the system identifies areas for improvement and generates patient-specific recommendations for lifestyle modification. This function supports the physician's tasks of problem identification and planning. For the TC function, the system inquires about the patient's actual condition and offers advice to the patient by interactive communication devices, i.e. mobile phones, personal digital assistances (PDA), and the Internet. This function supports the patient's task of implementation of the plan. The two functions work together to enhance adherence to lifestyle modifications and improve glycemic control in diabetic patients.

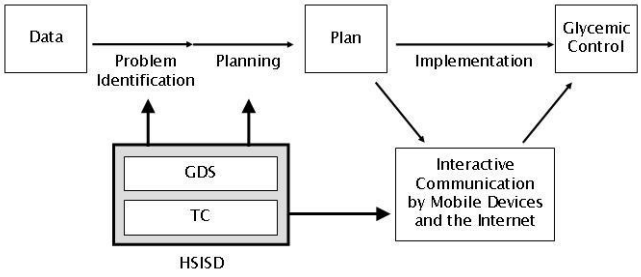

Fig.1. HSISD in Diabetes Management

HSISD is used in a hierarchical environment that consists of a knowledge control center, clinics and hospitals, and patients. The knowledge control center manages all the knowledge for GDS and TC. The guideline knowledgebase for GDS and the communication knowledgebase for TC are distributed to clinics and hospitals through the Internet. Each clinic and hospital has its own server connected to the Internet, and the requisite knowledge and data for GDS and TC are stored in it, to be used as needed.

Figure 2 summarizes the overall process of GDS and TC in a clinic or hospital. Practical functions of GDS and TC are implemented by the computer at the clinic or hospital.

The GDS function identifies areas for improvement and generates patient-specific recommendations for lifestyle modifications. Rule-based reasoning methodology is applied to the three steps of (1) identifying the current health behavior problems, (2) setting specific treatment goals, and (3) assigning priorities among the goals. The first step is based on diagnostic rules. The second and third steps are based on therapeutic rules. The diagnostic and therapeutic rules are stored in the guideline knowledgebase in the form of IF-THEN rules. Most of the rules are made based on knowledge gained from CPG and expert consensus.

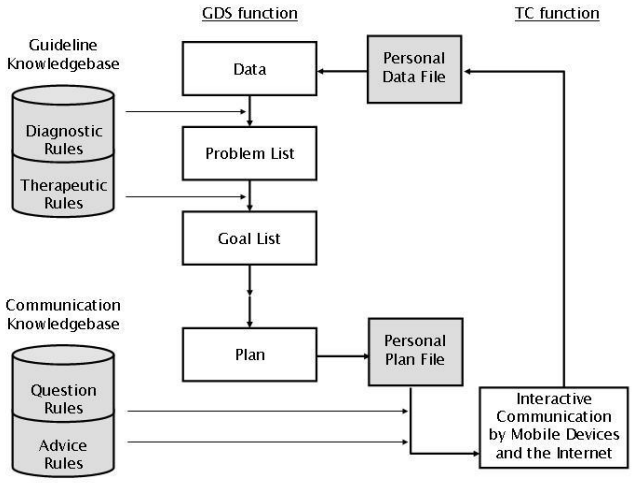

Fig.2. GDS and TC in a Clinic or Hospital

The TC function inquires about the patient's actual condition and offers advice to the patient by interactive communication devices. The patient receives an e-mail from the attending clinic or hospital with an invitation to answer a questionnaire on the linked Web page. The Web page is hosted and managed by the attending clinic or hospital. The questionnaire consists of two general questions, i.e. "How much do you weigh today?" and "Have you developed any symptoms of anxiety?" and two or three patient-specific questions. The patient-specific questions ask about the lifestyle behaviors targeted in the plan, e.g. "How often did you eat fried foods last week?" as shown in Fig. 3. As soon as the patient submits the answers, advice can be offered that will help in accomplishing the specific goals. Such TC is repeated almost once a week between the clinic or hospital visits using Cron. If the patient's condition becomes serious, however, as judged from the TC, the patient is urged to visit the clinic or hospital at an earlier date.

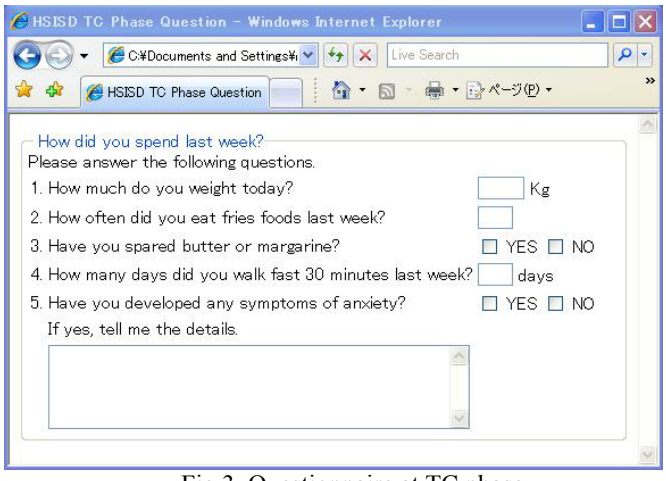

Fig.3. Questionnaire at TC phase

B. Self-Rating Depression Scale

When the patient answers a questionnaire at TC phase, the system also checks the depression rate of the patient. In this system, the depression of the patient is calculated.

1. I feel downhearted, blue, and sad.
2. Morning is when I feel the best.
3. I have crying spells or feel like it.
4. I have trouble sleeping through the night.
5. I eat as much as I used to.
6. I still enjoy sex.
7. I notice that I am losing weight.
8. I have trouble with constipation.
9. My heart beats faster than usual.
10. I get tired for no reason.
11. My mind is as clear as it used to be.
12. I find it easy to do the things I used to do.
13. I am restless and can't keep still.
14. I feel hopeful about the future.
15. I am more irritable than usual.
16. I find it easy to make decisions.
17. I feel that I am useful and needed.
18. My life is pretty full.
19. I feel that others would be better off if I were dead.
20. I still enjoy the things I used to do.

Fig.4. Check sheet for depression

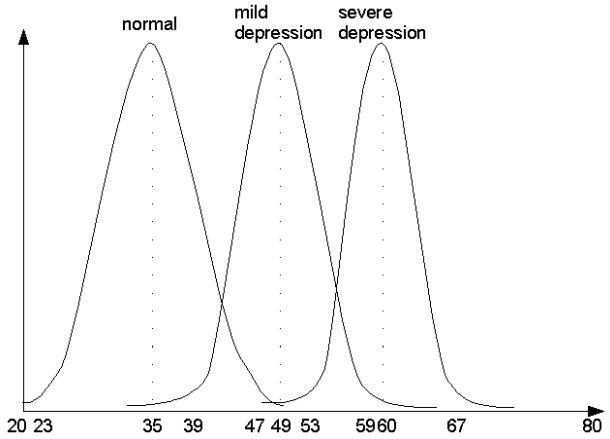

Fig.5 Fuzzy membership function for depression

In order to check the depression rate of the patient by himself/herself, we adopt "Zung Self-rating Depression Scale: SDS" [7]. It contains 20 items, with 10 items keyed negatively and 10 positively as shown in Fig. 4. For each item, the subject rates whether the item occurred 1 = "a little of the time," 2 = "some of the time," 3 = "a good part of the time," or 4 = "most of the time." To obtain a total severity score, positive items are reversed, and then all items are summed. SDS scores in this research are interpreted as follows: 23-47, within normal range; 39-59, minimal to mild depression; 53-67, moderate to severe depression as shown in Fig. 5. Our system recommends consulting a physician when the result is categorized into "moderate to severe depression."

III. Anxiety Type Classifying Method

In TC phase, most of the questions are filled by inputting the values and checking the boxes. However, the last question "Have you developed any symptoms of anxiety?" is free-answer style. The answers of the fix-style questions can be utilized to calculate the achievement of the patient's healthcare plan and the computer system can deal with such digital data easily. However, it is difficult to extract explicit information from raw text data like free-answer style description. Therefore, analyzing the type of anxiety topic of the patient in advance will support that the physician understands the text.

We propose a method to analyze free-answer style text data and classify them into topic types of patient's anxiety. Various anxiety descriptions about diabetes are collected in the Internet and are classified into the following five types of anxiety types; "mental problem," "physical problem," "diet," "physical activity," and "medicine." Mental problem contains anxiety to the future, fear about disease, depression, and so on. Physical problem contains complications, symptoms, result of blood test, and so on. Diet contains food, drink, glucose intake, and so on. Physical activity contains walking, daily activities, and so on. And medicine contains medicine intake, side effect, and so on.

Our proposed system can classify the input text data from patients into five patterns by using SOM which is trained by grammatical features in the text data.

A. Self Organizing Feature Map

The basic SOM [5] can be visualized as a sheet-like neural network array as shown in Fig. 6, the cells (or nodes) of which become specifically tuned to various input signal patterns or classes of patterns in an orderly fashion. The learning process is competitive and unsupervised, which means that no teacher is required to define the correct output for an input. Only one map node called a winner node at a time is activated corresponding to each input. The map consists of a regular grid of processing units. A model of some multidimensional observations, eventually a vector consisting of features, is associated with each unit. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models are far from each other.

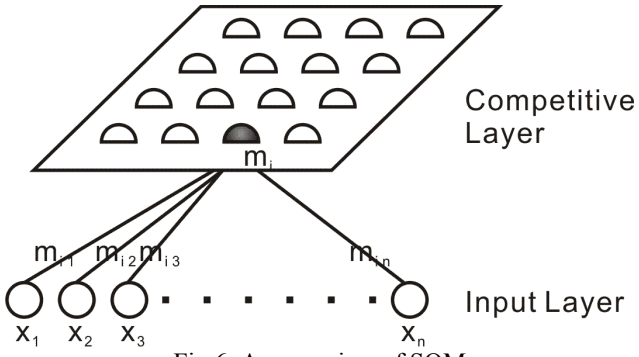

Fig.6. An overview of SOM

A sequential regression process usually carries out fitting to the model vectors. The n is the number of input signals. An input vector $\mathbf{x}$ is compared with all the model vectors $\mathbf{m}_i(t)$. The best-match unit on the map is identified. The unit is called the winner. For each sample $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, first the winner index c (best match) is identified by the condition.

$$\forall i, \|\mathbf{x} - \mathbf{m}_c\| \le \|\mathbf{x} - \mathbf{m}_i\| \quad (1)$$

After that, all model vectors or a subset of them that belong to nodes centered around node c are updated at time t as

$$\begin{aligned} &\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(\mathbf{x}(t) - \mathbf{m}_i(t)) && \text{for } \forall i \in N_c(t) \\ &\mathbf{m}_i(t+1) = \mathbf{m}_i(t) && \text{otherwise} \end{aligned} \quad (2)$$

Here h_{ci} is the neighborhood function, a decreasing function of the distance between the ith and cth nodes on map grid. The $N_c(t)$ specifies the neighborhood around the winner in the map array. This regression is usually reiterated over the available samples.

At the beginning of the learning process, the radius of the neighborhood is large and the range of radius becomes small according to the convergence state of learning. That is, as the radius gets smaller, the local correction of the model vectors in the map will be more specific. The h_{ci} also decreases during learning.

B. Training of SOM

In our proposed method, firstly, SOM is trained by the collected text of various free-answer style. Then, an input vector in SOM is generated from a text data, and the free-answer style text data is classified into five types of anxieties based on the label of the winner node. Each element in the input vector corresponds to the number of the valid words which can be classified into anxiety type as described in the following subsection 1). In order to pick up the valid words for classifying the topic types, the examples of the text data about diabetic problem are collected from some web-based bulletin board system and some relations between the frequency of appearance of the words and the topic types are found. Then we propose a method to transform the text data to the input vector.

1) Defining Feature Words

When an input text data is transformed to the input vector, we may note that the specified words as elements of the vector. First, such "feature words" should be selected.

We retrieved the text data about diabetic from the descriptions in some BBSs of network communities which contain the words such as "*Tounyoubyou* (diabetes)," "*Nayami* (annoyance)," "*Soudan* (counseling)," and so on.

The retrieved examples were analyzed morphologically and each word was transformed into the original form. Then, we calculated the frequencies of the words for each topic as;

$$F_n^{\,t} = \frac{n_t}{n_{all}} \quad (3)$$

$F_n^{\,t}$ indicates the frequency of the word n for topic t. n_t indicates the number of word n in the text data of topic t, and n_{all} indicates the number of word n in the all text data. The maximum value of $F_n^{\,t}$ was not 1.0 because $F_n^{\,t}$ is calculated by computer and the result contains round-off error. Therefore, we defined word n as feature word if the $F_n^{\,t}$ is more than 0.9995 intuitively. There are 126 feature words for anxiety. Table 1 shows some of the feature words.

TABLE I
FEATURE WORDS FOR ANXIETY (PARTLY)

Futoi (fat), *Inshurin* (insulin), *Karorii* (calorie), *Hasshou* (develop the symptom), *Shujutsu* (surgery), *Budoutou* (glucose), *Yuuutsu* (depression), *Jinzou* (kidney), *Bouin* (excessive drinking), *Kinniku* (muscle)

2) Generating Input Vector

In order to input the text data into the SOM, the text data should be transformed to the vector which consists of some elements. Each element shows the ratio of the number of the feature words in the sample data. The process to generate input vector from input data is described as follows. At first, the input text data is analyzed morphologically. Mecab [8] is the useful tool for Japanese morphological analyzer. Next, the appearance of each feature word in the input text data is counted. Word matching is implemented in the original form. The words in text data are compared with feature words. Then, an input vector consists of the agreement degree of feature words. At last, the pattern label is added to the vector by hand.

3) Map Training by SOM Algorithm

The trained map is constructed by using input data vectors as described above subsection 2). The SOM program can label the map units according to the input vectors. Each unit receives the labels (mental problem, physical problem, diet, physical activity, and medicine) of all the data vectors for which it is the best matching unit. The map units are then labeled according to the majority of labels "hitting" a particular map unit. The no "hit" units are left unlabeled.

Figure 7 shows a labeled map trained by using SOM with the input vectors. We can see that the symbols, "Mp4," "Pp1," and so on, indicated in the figure are the label names. "Mp" indicates "Mental Problem," "Pp" indicates "Physical Problem," "Di" indicates "Diet," "Pa" indicates "Physical Activity," and "Me" indicates "Medicine" and the following number is a serial number of input data. In this figure, we can see a cluster of "Pp" units at the left-bottom position. And there are a cluster of "Mp" and a cluster of "Pa" above the "Pp" cluster. At the right-middle position, there is a cluster of "Me."

Some units have multiple winner labels. A unit should have only one type of topic label because this map is used to output a topic type for an input. Therefore, when a unit has all the same type of topic labels or one type of label occupies the unit more than 50%, the unit is defined as the major label. When several types of labels share the node (i.e. no one label can occupy 50%), the label of the node becomes "unknown."

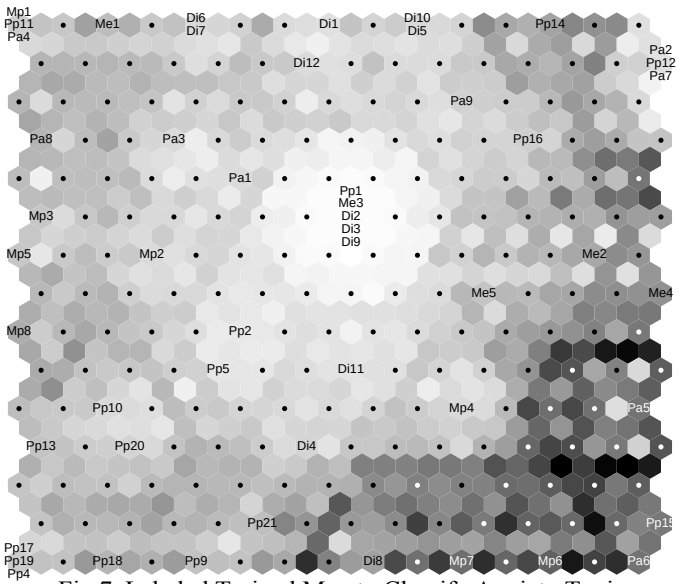

Fig.7. Labeled Trained Map to Classify Anxiety Topic

C. *Counseling Request Classifying Process*

In order to classify test data into five anxiety types, firstly, the data is analyzed morphologically and test data vector is generated in the same way. Secondly, the system applies the test data vector into the trained map and calculates the winner node. Then, the system finds the label of the winner node and outputs the label's name as the "classify result." One of the "mental problem," "physical problem," "diet," "physical activity," "medicine," "unknown," and "unlabeled" is output.

IV. Inner Emotion Classifying Method

We proposed a method to classify free-answer style text into five of anxiety types. When a physician checks the patient's statement, however, it is important not only "what the patient is anxious about" but also "how the patient feels about it." Therefore, we propose a method to analyze the patient's inner emotion by using SOM.

In this method, we classify the aroused emotion to the text into four emotions (Happy, Sad, Fear, and Anger). Firstly, we collected 38 learning data and defined 117 feature words as shown in Table 2.

The training data are transformed to input vectors and used to train SOM. Figure 8 shows a labeled map trained by using SOM with the input vectors. "Hp" indicates "Happy," "Sd" indicates "Sad," "Fe" indicates "Fear," and "Ag" indicates "Anger." In this figure, we can see a large cluster of "Sd" at the left side. There is "Fe" cluster at left-middle position and "Ag" cluster is at left-top position. There are little "Hp" labels at this map because most of learning data for "Happy" contained little feature words so generated input vectors were almost **0**. **0** are gathered at center position.

TABLE II
Feature Words for Emotion (partly)

Kujou (complaint), *Yuuutsu* (depression), *Zeitaku* (luxury), *Ikari* (angry), *Baka* (idiot), *Shosen* (after all), *Muri* (impossible), *Suki* (like), *Kirei* (beautiful), *Zankoku* (cruel)

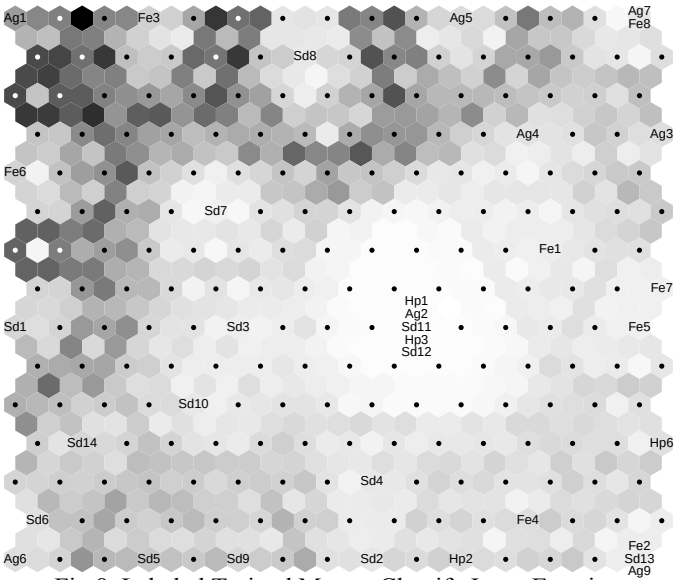

Fig.8. Labeled Trained Map to Classify Inner Emotion

V. Experimentation

We experimented with our method using a 15*15 sized map and it was trained by 5,000,000 epochs. Table 3 shows the experimental result for classification of anxious type. 55 data (Mental Problem: 8, Physical Problem: 21, Diet: 12, Physical Activity: 9, and Medicine: 5) were tested and 42 data (Mental Problem: 7, Physical Problem: 14, Diet: 9, Physical Activity: 8, and Medicine: 4) were classified correctly. 9 data outputted the center node because these input vectors were **0**. It indicates that the input vector transformed from text does not express its feature explicitly. If we increase feature words or extract another grammatical features, these data will be distinguished.

TABLE III
Experimental Result for Anxious Type

Type	Mp	Pp	Di	Pa	Me
Result	7/8	14/21	9/12	8/9	4/5

Table 4 is the experimental result for classification of inner emotion type. 38 data were tested and 24 data were classified correctly. 9 data outputted the center node because these input

vectors were **0**. Especially most of test data about "Happy" were classified into "unknown" because the SOM did not learn enough "Happy" data as described in Section 4. In order that the SOM learns more "Happy" data, the effective input vectors should be calculated by defining new adequate feature words.

TABLE IV
EXPERIMENTAL RESULT FOR INNER EMOTION

Type	Hp	Sd	Fe	Ag
Result	2/7	11/14	6/8	5/9

The output format of the result of our proposed method should be explicitly visualized for physicians because they usually treat patients considering only the physical symptoms and they do not have enough time to take into account of the mental problems. However, the mental aspect is very important to encourage and maintain the motivation of life style modification for diabetes treatment. Therefore, our system visualizes the analytical results of the patient's free-answer style descriptions as a radar chart as shown in Fig. 9. The number of appearance of each topic with negative emotion is counted for last 15 comments. The physicians can find out that what topic the patient feels displeasure for and they can ask a question about the topic. In the example of Fig. 9, physician will ask the patient like "How do you feel about the medicine intake recently?"

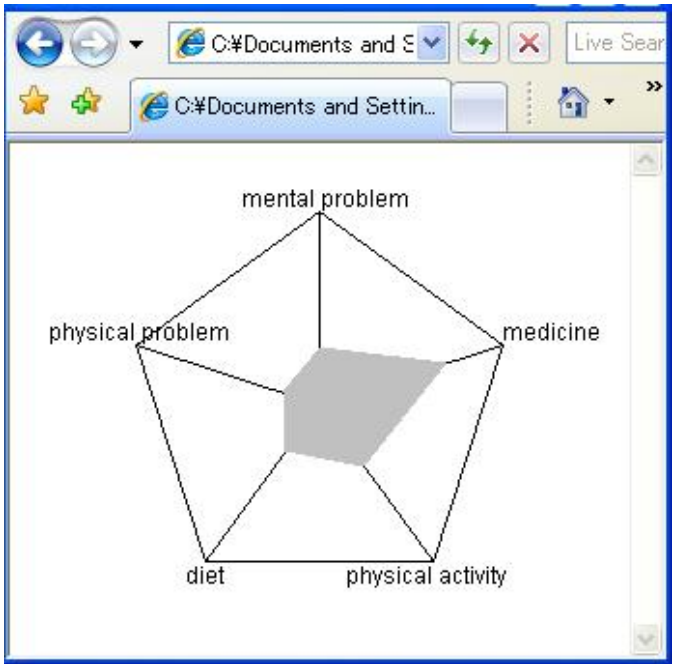

Fig.9. Visualized Analyze Result

VI. CONCLUSION

We have improved TC phase in the HSISD system. In order to support to understand the patient's appeal and to propose an adaptive healthcare plan to the patient, we proposed a method to analyze the patient's anxiety type (mental problem, physical problem, diet, physical activity, and medicine) and inner emotion (happy, sad, fear, and anger) from the free-answer style text using SOM. Firstly, we defined feature words to transform input text data to the vector. The vector is generated by analyzing input text morphologically and counting the number of feature words which featured the five anxiety types. Then SOM was labeled based on the labels of the training data. When a unit has all the same type of annoying pattern labels or one type of label occupies the unit more than 50%, we give the unit the major label. When some type of labels share the node, the node has "unknown" label.

We experimented with our proposed method by using the trained and labeled map and the correct answers were obtained for 42 data about anxiety type and 24 data about inner emotion type.

We will develop our method to be able to add new feature words automatically from the failed test data in near future. Furthermore, because the concept of the feature word is variable at various situations, an adaptive structure learning method of SOM [9] will be applied to this research.

REFERENCES

[1] Susman J.L., Helseth L.D., "Reducing the complications of type II diabetes: a patient-centered approach," *Am. Fam. Physician*, vol. 56, pp. 471-480, 1997.

[2] Wheeler M.L., "Nutrition management and physical activity as treatment for diabetes," *Diabetes*, vol.26, pp. 857-868, 1999.

[3] Day J.L., "Diabetic patient education: determinants of success," *Diabetes Metab. Res. Rev*., vol.16, suppl. 1, pp. S70-S74, 2000.

[4] Machi Suka, Takumi Ichimura, and Katsumi Yoshida. A Knowledge-based Intelligent System for Diabetes Management, *in "Knowledge Based Intelligent System for Healthcare,"*, pp.271-302, Advanced Knowledge International, 2004.

[5] T. Kohonen, "Self-Organizing Maps (3rd Ed.), *Springer-Verlag*, 1997.

[6] Ikki Ohmukai, "Current Status and Future Perspectives of Social Networking Services," Journal of Information Processing Society of Japan, Vol.47, No.9, pp. 993-1000, 2006 (in Japanese).

[7] Kazuhiko Fukuda, and Shigeo Kobayashi, "SDS – Self Depression Scale-user's manual, *Sankyobo*, 1983 (in Japanese).

[8] Mecab (http://mecab.sourceforge.net/)

[9] Takumi Ichimura, Shinichi Oeda, Toshiyuki Yamashita, and Eiichiro Tazaki, "A Learning Method of Neural Network with Lattice Architecture," Journal of Japan Society for Fuzzy Theory and Systems, Vol.14, No.1, pp.28-42, 2002.