The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

On classification of interview sheets for ophthalmic examinations using self-organizing maps

Naotake Kamiura[1], Ayumu Saitoh[1], Teijiro Isokawa[1], Nobuyuki Matsui[1], and Hitoshi Tabuchi[2]

[1]Graduate School of Engineering, University of Hyogo, Himeji 671-2280, Japan
(Tel: 81-79-267-4918, Fax: 81-79-266-8868)
[2]Tsukazaki Hospital, Himeji 671-1227, Japan

[1]{kamiura, saitoh, isokawa, matsui}@eng.u-hyogo.ac.jp

Abstract: In this paper, a method of determining examination groups for new patients is presented, using self-organizing maps. Assuming that interview sheets are divided into four classes, the method copes with the examination determination as the classification of the sheets. The data are generated from handwriting sentences in the sheets. Some nouns in them are picked up as elements of the data. After map learning is complete, its neurons are labeled. The class of the sheet corresponding to the data to be checked is specified by the label of the winner neuron for the data. It is established that the multiple-map-based scheme achieves favorable classification accuracy.

Keywords: data classification, interview sheets, self-organizing maps, waiting time problems

1 INTRODUCTION

In Japan, long waiting time has been considered to be one of the major reasons that prevent chronic patients from going to hospitals. The condition of chronic patients keeps on worsening, and the fatal damage tends to suddenly befall to such patients. In Ji, Yanagawa, and Miyazaki [1], an event-driven network approach using queuing theory is presented to reduce the waiting time. The dispatching rules are suggested based on patients' expected visitation time and expected service time, and they are used to schedule the patients. The advisability of applying the rules is sometimes restricted by situations associated with the numbers of clinical testing equipments, medical doctors, nurses, medical technologists and so forth.

Managing the waiting time as part of the examination time is also a promising approach. Before consultation, a doctor first reads an interview sheet filled out by a new patient. The doctor then determines a set of examinations for the new patient, whereas the patient must wait for a while before the doctor make a determination. This medical protocol also prevents the doctor from allotting enough time to see the patient. The doctors are therefore anxious for the system that automatically checks the interview sheets and determines a set of the examinations.

In this paper, the determination of examination groups for new patients is presented, using self-organizing maps (SOM's). SOM's are applied as useful classification tools. In Kurosawa et al [2] and Ohtsuka et al [3], [4], SOM's have strong research interest as a means of expressing and/or processing clinical examination results. The proposed method assumes that each of the interview sheets filled out by new patients belongs to one of the four classes. It therefore copes with the examination determination as the classification of data corresponding to the sheets. An open source engine developed for Japanese language morphological analysis is applied to handwriting sentences in the sheets, and some nouns in them are picked up as elements of the data presented to the maps. After general SOM learning is complete, neurons in the maps are labeled. The class of the data to be checked is specified by the label of the winner neuron. The scheme using a single map for all of the patients and that dividing patients into five age groups and preparing a map for each of the groups are discussed. It is revealed that the latter scheme is useful in achieving favorable classification accuracy, compared with the former scheme.

2 PRELIMINARIES

A map consists of neurons, and a neuron has a reference vector with M element values if the M-dimensional data is presented to the map. General SOM learning is conducted using the following formulas.

$$NF(t) = r_0\left(1 - t/T\right), \tag{1}$$

$$\tau_i(t) = \tau_0\left(1 - t/T\right), \tag{2}$$

$$W_i(t) \leftarrow W_i(t) + \tau_i(t)\left(X^l(t) - W_i(t)\right). \tag{3}$$

$X^l(t)$ means the l-th training data. The neighborhood function around the winner neuron for the data presented at time t is defined by Eq. (1). $W_i(t)$ is the reference vector of the i-th neuron, C_i. If C_i is located inside the area specified by $NF(t)$, $W_i(t)$ is modified according to Eqs. (2)

and (3). $\tau_l(t)$ is the learning rate. T is the epoch number employed as the learning-termination condition.

This paper focuses on sentences handwritten in Japanese by new patients visiting the department of ophthalmology. They appear in interview sheets. It is probable that patients contracting different diseases undergo same examinations. According to common characteristics in terms of the examinations for the diseases, this paper divides examinations into the following four classes: Class 1 associated with fundus examinations, Class 2 associated with glaucoma tests, Class 3 associated with slit lamp tests, and Class 4 associated with oculomotor tests. Class 1 is strongly linked to cataract, diabetes, retinal disease, uveitis, and pediatric ophthalmology. Class 3 is strongly linked to corneal and/or conjunctival disease, lacrimal apparatus, and ametropia. A similar relationsip holds between Class 4 and each of strabismus, neuro-ophthalmologic disease, and trauma. Since a new patient undergoes the examinations belonging to one of the four classes, the interview sheets can also be divided into such four classes. The proposed method then copes with determination of examinations as classification of data corresponding to interview sheets with the four classes.

3 DATA GENERATION FROM INTERVIEW SHEETS

Handwriting sentences in interview sheets are first electronically registered from the keyboard. MeCab, which is the open source engine developed for Japanese language morphological analysis by Kudo, Yamamoto, and Matsumoto [5], is then applied to divide the words in them into some parts of speech. Several of the nouns are picked up per sentence by consulting the list of prohibited words. Table 1 shows examples of the prohibited words.

Table 1. Examples of prohibited words

Types of prohibited words	Examples
Numbers	1, 2, 0, 100, 36.5
Symbols associated with SI base units	m, mm, cm, kg
Geographical names	姫路/ Himeji, 佐用/ Sayoh
Words associated with time	週/ week, 金曜日/ Friday, 去年/ last year, 先月/ last month

The proposed method next generates a matrix. Let us assume that N interview sheets whose classes are perfectly known are available. Let d denote the total number of nouns picked up in the above manner. The matrix then has N rows and d columns. In other words, the nouns are assigned to the columns, and each of the sentences (i.e., the interview sheets) corresponds to a row. Frequencies of appearance are initially given as follows: if the p-th noun appears in the l-th sentence m_{lp} times, the element specified by the l-th row and p-th column is set to m_{lp}, where $1 \leq l \leq N$ and $1 \leq p \leq d$. Fig. 1 depicts an example of the first matrix. In the following discussions, m_{lp} means the value given to the element specified by the l-th row and p-th column.

Weighting element values adequately is useful in emphasizing the significance of the words corresponding to such values. The proposed method weights some values, based on the probability of words appearing. Let us assume that the p-th noun appear NA_p^q times in the set of registered sentences (i.e., the interview sheets) belonging to Class q, where $1 \leq p \leq d$ and $1 \leq q \leq 4$. In addition, let R_p^q denote the ratio of the number of the p-th nouns appearing in the sentences belonging to Class q, compared to the total number of the sentences belonging to Class q. Assuming that the latter number is denoted by NS^q, R_p^q is as follows.

$$R_p^{\,q} = NA_p^{\,q} \,/ NS^q . \tag{4}$$

It is considered that nouns with comparatively high R_p^q's are of importance in specifying attributes of Class q. The threshold value associated with R_p^q is set to 0.055 when the noun is checked whether its element values are weighted.

The p-th noun is defined as a powerful word, if the value calculated by Eq. (4) is more than 0.055 for Class k solely, where $k \in \{1, 2, 3, 4\}$. We then have $R_p^k \geq 0.055$ and $R_p^q < 0.055$ for any q, where $q \in \{1, 2, 3, 4\}$ and $k \neq q$. If Eq. (4) calculates the value more than 0.055 for two classes, the noun is defined as a special word. In the case where the p-th noun is the special word, $R_p^q < 0.055$, $R_p^{k1} \geq 0.055$ and $R_p^{k2} \geq 0.055$ hold for $k1$, $k2$, and any q, where $k1, k2 \in \{1, 2, 3, 4\}$, $k1 \neq k2$, $k1 \neq q$, and $k2 \neq q$. The first matrix is modified as follows. Each of the element values on the columns corresponding to the powerful (or special) words is multiplied by 4 (or 2). For example, if Word 2 in Fig. 1 is the powerful word, element values on the second column are weighted and change from $(m_{12}, m_{22}, m_{32}, \ldots, m_{N2})=(0, 2, 0, \ldots, 1)$ to $(0, 8, 0, \ldots, 4)$. Besides, if Word 3 is the special word, we have $(m_{13}, m_{23}, m_{33}, \ldots, m_{N3})=(4, 0, 2, \ldots, 2)$ as weighted values.

It is difficult to specify the attribute of a unique class, using a noun that equally appears in the numerous sentences belonging to arbitrary classes. This is why targets of weighting are restricted to the nouns, each of which has the value calculated by Eq. (4) exceeding the threshold for at most two classes.

	Word 1	Word 2	Word 3	…	Word d
	検査 “Examination”	白内障 “Cataract”	眼鏡 “Glasses”	…	網膜 “Retina”
Sheet 1	0	0	2	…	1
Sheet 2	1	2	0	…	0
Sheet 3	1	0	1	…	0
⋮	⋮	⋮	⋮	⋱	⋮
Sheet N	0	1	1	…	0

Fig. 1. Example of first matrix

4 EXAMINATION DETERMINATION USING SELF-ORGANAIZING MAPS

For the use of SOM’s, two schemes are available. The first scheme is based on using a single map, whereas the other prepares a map for each age group. The maps are constructed in the general manner using Eqs. (1)-(3). The general manner is also adopted in Ohtsuka et al [3], [4]. Note that a row in a matrix generated by the method in Sect. 3 is presented to a map as a member of the training data set.

Let us discuss the single-map-based determination. The information associated with age is added as the rightmost (i.e., the (d+1)-th) column to the matrix. One of the three-level values is given to each element in the (d+1)-th column. If the patient filling out the l-th sheet is younger than 11 years old, 0 is given as m_{ld+1}. If the age of the patient is more than 10 and less than 46, we have m_{ld+1}=6. If it is larger than or equal to 46, m_{ld+1}=12 holds.

Once general SOM learning is complete, neurons in the map are labeled as follows.

<Neuron labeling>

[Step 1] Let F_q^i denote the frequency of the i-th neuron (C_i) firing for the training data belonging to Class q, where $1 \leq q \leq 4$. Set four F_q^i’s to 0, and set l to 1.

[Step 2] The l-th training data is presented, and F_q^i’s of the winner are updated. The value of l is incremented by 1.

[Step 3] If $l \leq N$, go to Step 2; otherwise, go to Step 4. Note that N is the total number of training data.

[Step 4] Let LN^i denote the label of C_i. It is as follows.

$$LN^i = \arg\left\{\max_q \left(F_q^i\right)\right\}. \tag{5}$$

Labels are assigned to the other neurons, using Eq. (5).

In this paper, each of the data unused for SOM learning is referred to as pilot data. The set of pilot data is also generated in the manner described in Sect.3. When one of the pilot data is examined, it is presented to the map with labeled neurons. The class of the data is considered to be that specified by the label of the winner for it.

Let us next explain the determination employing multiple maps. The following five age groups are defined: Group 1 with patients whose age is less than 31, Group 2 with patients whose age is more than 30 and less than 56, Group 3 with patients whose age is more than 55 and less than 66, Group 4 with patients whose age is more than 65 and less than 76, and Group 5 with patients whose age is more than 75. A training data set with four classes is generated for each group, and a map intended for exclusive use with a group is constructed by using it. The (d+1)-th column associated with age is ignored for learning, labeling, and data classification. The above labeling method is also adopted. Classifying the pilot data is also similar to that applied for the first scheme, except that the map used depends on age of the patient corresponding to the data.

5 EXPERIMENTAL RESULTS

The proposed method was applied to interview sheets provided from Tsukazaki hospital in Japan. A map with ten rows and ten columns is prepared. It is trained, subject to T=1000 in Eqs. (1) and (2). Besides, r_0=20 and τ_0=1.0 hold for initial values of in Eqs. (1) and (2).

Let IS_q^k denote the number of pilot data, each of which is judged as Class q while its actual class is Class k, where k, $q \in \{1, 2, 3, 4\}$. The percentage of the number of pilot data whose classes are judged as Class q, compared to the total number of pilot data actually belonging to Class k, is calculated. The following value is referred to as the percentage of concordance associated with Class k, PC_k.

$$PC_k = 100 \times IS_k^k \,/\, \left(\sum_{q=1}^{4} IS_q^k\right). \tag{6}$$

The experimental results for the scheme using a single map are first shown. A set of data generated from interview sheets for 580 patients is used. The sheets were filled out from May through November 2010. The correspondence between a sheet and its class is perfectly known in advance. The sheets for hundred patients are randomly chosen, and a set of pilot data is generated from them. A set of training data is then generated from the remaining sheets, and a map is constructed by presenting members in the training data set. The classification capability of the resultant map is evaluated, using the pilot data set. A trial consisting of the above is repeated ten times, and mean values of PC_k's are obtained. The results are tabulated in Table 2. Note that unfavorable values associated with the percentages of discordance also appear in it. We have PC_1=82.0, PC_2=48.8, PC_3=62.8, and PC_4=48.8. PC_1 is the highest value of four PC_k's, and the proposed method achieves favorable PC_3. Both of PC_2 and PC_4, however, are less than fifty percent. This is due to the fact that the data actually belonging either to Class 2 or to Class 4 tend to be misjudged as the data of Class 1. Inhibiting misjudgment from Class 2 (or 4) to Class 1 is especially crucial in improving capability of the single-map-based classification.

Let us next show the experimental results for the five-map-based classification. The number of interview sheets is 2407. They were filled out from May through November 2010. The classes of data generated from them are perfectly known in advance. The data of fifty patients are randomly chosen as pilot data for each of the age groups (Groups 1 through 5). The remaining data are used as training data. Recall that a map is constructed for each group. The five-map-based method is evaluated by classifying the pilot data. A trial consisting of the above is repeated ten times, and the following value, PC_{ave}, is calculated for each group from averaged PC_k's.

$$PC_{ave} = \left(\sum_{q=1}^{4} PC_q\right)/4. \qquad (8)$$

PC_{ave}'s are tabulated in Table 3. Patients in Group 3 suffer from a wide range of diseases compared with patients in any other group. This is why PC_{ave} is somewhat disappointing for Group 3. The method, however, achieves favorable PC_{ave}'s for other groups. It is thus established that the classification should be conducted for each age group.

Table 2. Classification results achieved by single-map-based method

		Classification results (%)			
		Class 1	Class 2	Class 3	Class 4
Actualities	Class 1	82.0	4.0	12.0	2.0
	Class 2	42.4	48.8	8.4	0.4
	Class 3	21.6	0.0	62.8	15.6
	Class 4	36.8	1.6	12.8	48.8

Table 3. Classification results achieved by five-map-based method

Age groups	PC_{ave}'s (%)
Group 1	75.2
Group 2	65.2
Group 3	59.6
Group 4	72.2
Group 5	78.0

6 CONCLUSIONS

This paper proposed the SOM-based method of determining examination groups for new patients from their interview sheets. MeCab is applied to handwriting sentences in the sheets, and several of the nouns are picked up per sentence. A matrix in which the sheets (or nouns) are related with the rows (or columns) is then generated. Frequencies of the nouns appearing in the sentences are basically given as element values in the matrix, and its row is presented to a map as a member of a training data set. The proposed method determines the examination group for some patient by classifying the data generated from the interview sheet filled out by the patient. The class of data is specified by the label assigned to the winner neuron for the data. From experimental results, it has been revealed that the scheme using maps for five age groups is superior in classification accuracy to the single-map-based scheme.

In future studies, the proposed method will be modified to improve the classification accuracy.

REFERENCES

[1] Ji Y, Yanagawa Y, and Miyazaki S (2010), Reducing outpatient waiting times for hospital using queuing theory (in Japanese). Journal of Japan Industrial Management Association, Vol. 60, No. 6, Feb., pp. 297-305

[2] Kurosawa H, Maniwa Y, Fujimura K, Tokutaka H, and Ohkita M (2003), Construction of checkup system by self-organizing maps, Proc. of Workshop on Self-Organizeing Maps, Kitakyushu, Japan, Sept.11-14, pp.144-149

[3] Ohtsuka A, Kamiura N, Isokawa T, Okamoto M, Koeda N, and Matsui N (2005), A self-organizing map approach for detecting confusion between block samples, SICE Trans., vol.41, no.7, July, pp.587-595

[4] Ohtsuka A, Tanii H, Kamiura N, Isokawa T, and Matsui N (2007), Self-organizing map based data detection of hematopoietic tumors, IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences, vol. E90-A, no.6, June, pp.1170-1179

[5] Kudo T, Yamamoto K, and Matsumoto Y (2004), Applying conditional random fields to Japanese morphological analysis, Proc. of Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, July 25-26, pp. 230-237