Classification for Data of Hematopoietic Tumor Patients with Fast Block - Matching - Based Self - Organizing Map Learning in Dynamic Environments

Naotake KAMIURA *¹ • Hirotugu TANII *² • Akitsugu OHTSUKA *³ •
Teijiro ISOKAWA *¹ • Nobuyuki MATSUI *¹

In this paper, a scheme of recognizing hematopoietic tumor patients is presented, using self-organizing maps constructed by fast block-matching-based learning. This fast learning is referred to as T-BMSOM learning. To classify the patients, screening data of examinees are presented to a constructed map. In T-BMSOM learning, a set of neurons arranged in square is regarded as a block, and one of the blocks is chosen as a winner per the presented data. It is assumed that members of a training data set to construct the map never change in static environments, whereas the data set is suddenly updated during learning in dynamic environments. While adopting the concept of blocks makes it possible to construct well-organized maps in dynamic environments, it lengthens the time for learning. To overcome this issue, T-BMSOM learning is based on a decision-tree-like winner search and a batch process. The screening data of an examinee frequently lacks several of the item values, and hence the data is presented to the map after averages of non-missing item values substitute for items with no values. The class of the data to be classified is basically judged by observing the label of a winner block. Simulation results establish that the proposed scheme achieves high accuracy of correctly recognizing the data of hematopoietic tumor patients, even if training the map is conducted in a dynamic environment.

Keywords : Self-organizing map, Block-matching-based learning, Dynamic environment, Hematopoietic tumors, Data classification

1. Introduction

Chemical therapy, radiation therapy and molecular-targeted therapy are medical cures for hematopoietic tumors such as leukemia, lymphoma and multiple myeloma. While these cures are advancing by quantum leaps, early detection remains the first line of defense against the tumors. In the urban area, sick people usually go to major medical institutions. Then, blood physicians, who are specialists in hematopoietic tumors, can directly examine them for signs of the tumors. In provinces where the number of blood physicians is limited, mostly sick people consult their home doctors first. If the home doctors are unfortunately inexperienced, it is difficult for them to properly diag-

nose the tumors. Both blood physicians and home doctors are therefore anxious for a system detecting the tumors at the early stage. The detection system will drastically shorten time wasted before effective treatment begins.

On the other hand, self-organizing maps (SOM's) [1]-[4] have recently attracted strong research interest as a means of expressing clinical examination results [5]-[7]. SOM's have the capability of projecting high-dimensional input data space onto a lower-dimensional map, preserving spatial or topological relationships in training data. In other words, if a training data set is well generated from examination results and a map is constructed using it, some neuron clusters are clearly formed in the map, according to attributes of data. The examination results visualized in such a way are intuitively comprehensible for general examinees and/or inexperienced home doctors,

*1 Graduate School of Engineering, University of Hyogo

*2 Fujitsu Ten Ltd.

*3 Yamato Scale Co., Ltd.

compared with the results expressed in terms of complicated lists of numbers. This especially applies to the case where each of the examination results consists of a large number of items. In [8]–[10], SOM's are adopted as tools to recognize the data of hematopoietic tumor patients.

A map is trained either in a static environment or in a dynamic environment. If a training data set is updated during learning in terms of numbers of its members, items and so on, it is said that a map is in the dynamic environment; otherwise, a map is in the static environment. Conventional SOM learning is controlled by two parameters referred to as neighborhood function and learning rate. They take small values as time advances during learning, and make it possible to construct a well-organized map in the static environment. Meanwhile, in the dynamic environment, these parameters often work inappropriately, and hence employing conventional learning results in a map with the limited capability in visually classifying the data. In [11], a set of neurons in the map is regarded as a block, and finding a winner is conducted at the block level during learning. Since this block-matching-based SOM (BMSOM) learning is never controlled by parameters monotonously decreasing with time, it yields the well-organized map even if environmental changes arise during learning. Adopting the concept of blocks, however, brings about marked increase in computational time complexity for learning.

In this paper, fast learning referred to as the T-BMSOM is conducted. To reduce computational time complexity, T-BMSOM learning employs the concepts of a decision-tree-like winner search and a batch process. The decision-tree-like search systematically narrows the region where a winner is located. T-BMSOM learning therefore requires fewer frequency of calculating the Euclidean distance per the training data than ordinary BMSOM learning in [11]. The batch learning process accumulates degrees and ratios of modifications on neuron reference vectors until the last member in the training data set is presented, and updates all of the vectors once an epoch. Due to this process, T-BMSOM learning outperforms ordinary BMSOM learning in terms of computational complexity relative to recalculating reference vectors at the block level.

This paper focuses on screening data of examinees, and proposes a scheme of recognizing hematopoietic tumor patients data, using maps constructed by T-BMSOM learning. The screening data of each examinee frequently lacks several of the item values. Therefore, before the data is presented to a map, it is processed as follows: the average of non-missing item values substitutes for each of corresponding items with no values of the above incomplete screening data. A block consists of neurons arranged in square. In T-BMSOM learning, the winner search is executed, examining each block in terms of the Euclidean distance. Once T-BMSOM learning is complete, every block is labeled, according to the Euclidean distance to the average of training data that belong to each of the following classes: normal persons, hematopoietic tumor patients, and non-hematopoietic tumor patients. In other words, the label assigned to the block means the class of averaged data with the closest distance. In classification phase as well as in learning phase, a block is also chosen as a winner, each time the data to be classified is presented to the constructed map. The class that the presented data belongs to is basically judged by observing the label of the winner block.

Provided that environmental changes in the training data set arise during learning, T-BMSOM-based classification is simulated. As a result, the data prepared for evaluation are categorized into the above three classes with a high degree of accuracy. In addition, examples of actually constructed maps are given. They serve to illustrate that clear clusters with equal-sized blocks are formed. This implies that the constructed maps can visually report dependable levels of classification results to users.

2. Preliminaries

This paper focuses on screening data that blood physicians actually use to diagnose hematopoietic tumors. They are results of blood and biochemical tests. The classification system should be designed so that the home doctors can use it easily. Since the screening data are easily available for home doctors, it seems that the screening data is one of the most promising examination results to be entered into the system.

The screening data consists of 56 items shown in Table 1. Note that examinee's age is added as the first

item. In this paper, the data basically belongs to one of the following three classes : normal persons (hereafter CN), hematopoietic tumor patients (hereafter CHT), and non-hematopoietic tumor patients (hereafter CNH).

In [8]–[10], a classification phase follows a learning phase. In the learning phase, a training data set is first generated from an original data set. The conventional algorithm then constructs maps, presenting its members. Conventional SOM learning absolutely requires the neighborhood function and learning rate, which take small values with time. Let us assume that environmental changes such as the updating of the training data set arise as shown in Fig.1. In the urban area where specialized users such as blood physicians work, the influence of such changes is not serious. This is due to the fact that the specialized users can adjust the above two time-controlled parameters from the outside of maps in the learning phase. In other words, the specialized users restart conventional SOM learning after making values of the parameters large, every time the changes arise. In this context, conventional SOM learning probably constructs well-ordered maps.

The specialized users do not always work full-time in the provinces. The classification system should then be designed so that it can be robust against unfore-

seen circumstances. The case where environmental changes are undetectable in the learning phase corresponds to one of such circumstances. In this case, the time for the user intervention from the outside of the map can never be specified, and hence conventional learning would result in inappropriate maps (e.g., twisted maps). In addition, it is possible that the intervention is even inapplicable for some reason. The probability that conventional learning constructs well-organized maps will further decrease in such a doubly awkward circumstance.

To overcome the above issue of environmental changes, the concept of blocks is introduced in [11]. In this ordinary BMSOM learning, finding a winner is carried out for each square block as shown in Fig.2. If the size of a map is N rows $\times$ N columns, the maximum (or minimum) size of a block is $N \times N$ (or 2×2). The block also has a reference vector. It equals the average of reference vectors of neurons in the block. Ordinary BMSOM learning examines blocks in terms of the Euclidean distance between the reference vector of each block and the presented training data. It then chooses the block with the shortest distance relative to the presented data as a winner. The reference vectors of neurons in the winner block are updated, each time the training data is presented. The updating of such neuron reference vectors depends on the inverse of the number of neurons in the winner (hereafter the winner size).

The block and inverse of winner size seemingly resemble the neighborhood function and learning rate in the conventional SOM, respectively. The size of the block to be chosen as a winner is never fixed to a certain value by user intervention before finding the winner. Determining the winner size and its inverse only

Table 1 Items of screening data

Item number	Abbreviation	Item
1	AGE	Age of an examinee
2	WBC	White blood cells
3	RBC	Red blood cells
4	HGB	Hemoglobin
⋮	⋮	⋮
⋮	⋮	⋮
54	NA	Sodium
55	K	Potassium
56	CL	Chlorine

Fig.1 Learning phase in dynamic environment

Fig.2 Block and map

depends on the reference vectors when the training data is presented. In other words, nothing is calculated and employed to determine the above two parameters, except for the Euclidean distance between the presented training data and each of the reference vectors. Since this characteristic also applies to the moments of the environmental changes shown in Fig. 1, the two parameters do not always decrease monotonously as time advances. Blocks thus work as the mechanism to manage the environmental changes without user intervention from the outside.

3. Fast Block-Matching-Based Learning

Ordinary BMSOM's proposed in [11] require more computational time complexity for learning than conventional SOM's. This is due to the fact that the number of all of the possible blocks is $(2N^3 - 3N^2 + N) / 6$ for an $N \times N$ -sized map. In ordinary BMSOM learning, all of the blocks are examined to find a winner. The frequency of calculating the Euclidean distance is therefore much larger than N^2 , which is equal to the calculation frequency required for conventional SOM learning. To cope well with this issue, the following two concepts, a decision-tree-like search of winner and a batch process, are adopted.

3.1 Decision-Tree-Like Search of Winner

Let us consider three blocks A, B and C in a map. Blocks A, B and C have s^2 , $(s-1)^2$, and s^2 neurons, respectively. Let us assume that block B is included in block A, and never overlaps with block C, as shown in Fig.3. In addition, block B is chosen as a winner for some training data. Block A then tends to be close to block B in terms of attributes of the block reference vectors to some extent. Besides, the Euclidean dis-

Fig.3 Winner block B and other two blocks

tance between the reference vector of block C and the presented data is probably longer than that associated with block A. The decision-tree-like search is based on the above intuitions. It is summarized as follows.

The decision-tree-like search chooses $(N-2)$ candidates for winner, each time a member of the training data set is presented to an $N \times N$ -sized map. The candidates differ from one another in sizes. Since the map consists of four $(N-1) \times (N-1)$ -sized blocks, they are examined first. The Euclidean distance is calculated between the reference vector and the presented data for each of such blocks, and one of them is determined as the first candidate. Thus, the decision-tree-like search is conducted on the assumption that the maximum block size is $(N-1) \times (N-1)$.

The remaining $(N-3)$ candidates are determined in the similar way. In other words, the $(N+1-s)$ -th candidate is determined out of the four $(s-1) \times (s-1)$ -sized blocks included in the same $s \times s$ -sized block, which has been most recently determined as the $(N-s)$ -th candidate. The smallest candidate is the $(N-2)$ -th candidate, and its size is 2×2 . The block with the shortest Euclidean distance is finally chosen as the winner for the presented data, out of such $(N-2)$ candidates. Fig.4 depicts an example of the search for a 5×5 -sized map.

Let us consider the frequency of calculating the Euclidean distance per member of the training data set. The decision-tree-like search examines four blocks to determine one candidate, and the number

Fig.4 Example of decision-tree-like winner search

of candidates is $N-2$. We therefore have $4(N-2)$ as the calculation frequency per member. Note that it is smaller than N^2 equal to the frequency required for conventional SOM learning. The decision-tree-like winner search is thus useful in reducing part of computational time complexity.

3.2 Batch Learning Process

In ordinary BMSOM learning, the recalculation of block reference vectors accompanies the updating of neuron reference vectors. Each time the training data is presented, it must be conducted for arbitrary blocks to reflect changes of neuron reference vectors in blocks. The batch learning process seems to be the promising way of reducing part of computational time complexity relative to the above recalculation.

In the batch learning process, numerical values associated with the modifications on neuron reference vectors are stored for every member in the training data set. Once an epoch is complete, all neuron reference vectors are updated at once. In other words, the recalculation of block reference vectors is invoked one time per epoch. To adopt this process, the following formulas are defined for the reference vector of the i -th neuron.

$$W_{update}^i(tc, j) \leftarrow W_{update}^i(tc, j-1) + \frac{x(tc, j)}{\alpha}, \quad (1)$$

$$W_{ratio}^i(tc, j) \leftarrow W_{ratio}^i(tc, j-1) + \frac{1}{\alpha}, \quad (2)$$

where $x(tc, j)$ is the j -th member in the training data set on the tc -th epoch, and $1 \leq tc \leq T$. α is the size of the winner block including the i -th neuron. Eqs. (1) and (2) are related to a degree and ratio of the modification on the i -th neuron. Thus, every time the training data is presented, the values obtained by Eqs. (1) and (2) are accumulated. When the tc -th epoch is complete, the reference vector of the i -th neuron, which is denoted by $W^i(tc)$, is as follows :

$$W^i(tc) = \frac{W_{update}^i(tc, Q)}{W_{ratio}^i(tc, Q)}, \quad (3)$$

where Q is the total number of members in the training data set.

The batch learning process is combined with the decision-tree-like winner search. Resultant fast learn-

ing is referred to as T-BMSOM learning. It is summarized as follows.

<T-BMSOM learning>

Step 1 All neuron reference vectors, tc and j are initially set to small random values, 1 and 1, respectively. All block reference vectors are then calculated.

Step 2 A winner block is determined for the j -th member, $x(tc, j)$, using the decision-tree-like search. $W_{update}^i(tc, j)$'s and $W_{ratio}^i(tc, j)$'s are calculated for all of the neurons in the winner.

Step 3 Set j to $j+1$. If $j < Q$, go to *Step 2* ; otherwise, set j to 1 and go to *Step 4*.

Step 4 $W^i(tc)$'s are calculated for arbitrary neurons. All block reference vectors are then calculated. Set tc to $tc+1$. The completion condition is specified by the maximum number of epochs. If it is fulfilled, stop learning ; otherwise, go to *Step 2*.

4 . Classification of Screening Data for Hematopoietic Tumor Patients

4.1 Generation of Training Data

Each time blood and biochemical tests are made to acquire the screening data, blood physicians instruct inspection officers which item values to measure. Various instructions then emanate from different physicians. In addition, item values possible to measure depend on the performance of examination devices. Medical institutions are not always furnished with identical devices. From the above, the data of an examinee frequently lacks several of the item values. Training a map using such incomplete data tends to degrade its capability of distinguishing classes, CN, CHT and CNH. The following imputation is therefore made for each of the data with missing item values.

A set of the screening data of which the classes are rightly known is prepared. Let us assume that the number of data, which belong to some class, is $P+U$, and P data out of them lack values of the k -th item, where $1 \leq k \leq 56$. Then, for the k -th item, the average of the remaining U data is calculated, and it is given to each of the above P data as the k -th item value. Fig.5 depicts an example of this imputation. Note that averaged values are also given to the h -th and e -th items. After the imputation is complete for all of the missing items, the set of 56-dimensional data is applied as a training data set.

Fig.5 Substitution carried out for generation of training data belonging to CHT

4.2 Labeling of Blocks

A label is assigned to a block as follows. First, three data are generated from all of the members in the training data set. For all of the items in the data, averages of members belonging to CN are calculated. The data relative to CN is thus generated. The remaining two data relative to CNH and CHT are similarly generated. Let ALD_r denote the data generated in the above way, using the training data belonging to the class r , where $r \in \{CN, CNH, CHT\}$. To assign a label to a block B, the Euclidean distance is calculated between the reference vector of B and each of ALD_r 's. The label of B equals that of the data with the minimum distance to B. For example, let us assume the case where the distance between B and ALD_{CN} , that between B and ALD_{CNH} , and that between B and ALD_{CHT} are 5, 15 and 10, respectively. B is then labeled "CN." If the data is equal to other data in distance to B, B is labeled "SUS." For example, if the distances to B are 5, 15 and 5 for ALD_{CN} , ALD_{CNH} and ALD_{CHT} , respectively, SUSPENSION (SUS for short) is given to B as its label.

4.3 Judgment Using Maps

To classify the data with some missing item values, three data are generated from each of the original data. For arbitrary items with no values, the averages of the training data belonging to each of CN, CHT and CNH are given to the original data as the corresponding item values. Fig.6 illustrates this generation. Note that three generated data also have labels. Each of the labels means the class of the training data used to calculate the averages in the above way.

The three generated data are presented to a trained map with labeled blocks. If the label of the generated

Fig.6 Generation of three data used for classification

data is different from the label of a winner block, a response, which is obtained by means of presenting it, is disregarded. Let us assume that the label of one of the three generated data only equals the label of a winner block. In this case, the number of available responses is one, and hence it is considered that the original data belongs to the class denoted by the latter label. Table 2 (a) shows an example. The class of the original data is considered to be CHT.

Let us next assume that labels of the two generated data at least equal labels of winners. In this case, the number of available responses is larger than one. To classify the original data, the Euclidean distance is checked between the reference vector of each winner and each of the presented data corresponding to winners. The class is considered to be that denoted by the label of the winner with the shortest distance. Table 2 (b) shows an example. It is presumed that the class of the original data is CN, because the response caused by the data with CN label has the shortest distance of the three available responses. Finally, if there are no available responses, the class of the original data is considered to be SUS. This means that judging the class is suspended.

In the learning phase including labeling of blocks and the classification phase, a winner is simply determined, based on the Euclidean distance between the presented data and the block reference vector. It seems that determining winners in this way is the comparatively careless classification with limited capability of accurately reflecting conditions of patients. There are some items for which normal ranges of possible values are clearly known. It is often important to observe whether measured item values deviate in the positive directions or in the negative directions from

Table 2 Example of judgments
(a) Case of single available response

		Labels of winners			Distance
		CN	CNH	CHT	
Labels of data	CN		√		15
	CNH	√			5
	CHT			√	10

(b) Case of three available responses

		Labels of winners			Distance
		CN	CNH	CHT	
Labels of data	CN	√			5
	CNH		√		15
	CHT			√	10

the normal range. If it is possible to realize powerful winner search schemes, which are devised according to knowledge on the normal ranges, deviating directions and so forth, the accuracy of recognizing the data of CHT will be improved.

It is, however, probably difficult to modify the proposed classification so that winners can be chosen in such a powerful way, assuming environmental changes occur during learning. For example, when new items are suddenly added to the training data as discussed in Subsects. 5.2 and 6.2, it would be hard to match up the new items with existing items during learning. In addition, the simple winner search scheme seems to be suitable, provided that inexperienced persons are included as users. The above are why the simple calculation of the Euclidean distance is executed for winner searches.

5. Classification Models in Dynamic Environments

The BM^S case hereafter means the case where the classification method in Sect.4 is applied in a static environment. On the other hand, the following two model cases are assumed to be dynamic environments.

5.1 Environmental Change in Number of Training Data

Let us focus on the model case where the number of training data changes with time. This case is hereafter referred to as the BM^{D1} case.

A set of screening data whose classes are known in advance is first divided into H subsets. The number of members in a subset is approximately equal to that

Fig.7 Environmental changes in number of training data

in another subset. The subsets are numbered, and members of the first through the M -th subsets are reserved for training a map. Members of the remaining $(H-M)$ subsets are employed as the pilot data, which means the data prepared to evaluate the proposed classification. A simulation consists of M stages. The training data set at the $(l-1)$ -th stage has the above-described members of the first through the $(l-1)$ -th subsets. Fig.7 illustrates the training data sets at the $(l-1)$ -th and l -th stages. Once the $(l-1)$ -th stage is complete, the training data set is updated by adding members of the l -th subset to it. T-BMSOM learning is then executed, provided that neuron reference vectors in a map are initially equal to those in the $(l-1)$ -th map. Note that the $(l-1)$ -th map means the map finally constructed at the $(l-1)$ -th stage. The pilot data are presented to the l -th map. When all of the pilot data are classified, the l -th stage is complete. The above are repeated for every stage.

5.2 Environmental Change in Number of Items

Let us next discuss the model case where the number of items used for training a map changes with time. This case is hereafter referred to as the BM^{D2} case.

First, in items are randomly chosen as members in the initial item set. The training data consists of values of such members, and hence it is seemingly regarded as the in -dimensional vector. T-BMSOM learning is conducted at the first stage. In each of the succeeding stages, the training data is updated by adding an item, which is randomly chosen out of the remaining items, to the data at the previous stage as a new element. A map is trained by presenting such updated data. Note that the substitution illustrated by Fig.5 is always executed for items without values. Fig.8

Fig.8 Environmental changes in number of items of training data

serves to explain this BM^{D2} case. The training data seemingly has $(in+q-1)$ items at the q -th stage. The q -th map is constructed, using such training data. Let us assume that the 56th item is chosen and added to the data at the $(q-1)$ -th stage. On condition that neuron reference vectors are initially equal to those in the q -th map, the $(q+1)$ -th map is trained with the updated $(in+q)$ -dimensional data. The above are repeated for every stage.

6. Experimental Results

The screening data is prepared for each of the 1831 examinees. Recall that their classes are known in advance. For the data of any examinee, about 28 percent of the 56 item values are missing on average. In the BM^S and BM^{D2} cases, all of the data are divided into five combinations by means of the five-fold cross-validation. A combination consists of 1465 training data and 366 pilot data.

The following percentage is assessed as one of the metrics for the classification capability: the percentage of the number of the data judged as belonging to one of the four classes (CN, CNH, CHT and SUS), compared to total number of the data belonging to each of the actual classes. Especially, the percentages of the numbers of data whose classes are precisely judged as CN, CNH and CHT are denoted by P^{CN} , P^{CNH} and P^{CHT} , respectively. For example, when the proposed method classifies the data as shown in Table 3, P^{CHT} is given by

$$P^{CHT} = \frac{n_2^{CHT}}{\sum_{d=0}^3 n_d^{CHT}}. \quad (4)$$

Note that each of the entries in Table 3 denotes the number of classified data.

Let us assume that a new class can be created, by

Table 3 Actualities versus classification results

		Number of classified original data			
		CN	CNH	CHT	SUS
Actualities	CN	n_0^{CN}	n_1^{CN}	n_2^{CN}	n_3^{CN}
	CNH	n_0^{CNH}	n_1^{CNH}	n_2^{CNH}	n_3^{CNH}
	CHT	n_0^{CHT}	n_1^{CHT}	n_2^{CHT}	n_3^{CHT}

combining CN, CNH and SUS. The following P_{sens}^1 and P_{spec}^1 are then defined as the first pair of sensitivity and specificity.

$$P_{sens}^1 = \frac{n_2^{CHT}}{\sum_{d=0}^3 n_d^{CHT}}, \quad (5)$$

$$P_{spec}^1 = 1 - \frac{n_2^{CN} + n_2^{CNH}}{\sum_{d=0}^3 n_d^{CN} + \sum_{d=0}^3 n_d^{CNH}}. \quad (6)$$

Note that P_{sens}^1 equals P^{CHT} .

If a combination of CNH, CHT and SUS is regarded as the suspicious class, another pair of sensitivity and specificity can be calculated. Let P_{sens}^2 and P_{spec}^2 denote such sensitivity and specificity, respectively. They are as follows.

$$P_{sens}^2 = \frac{\sum_{d=1}^3 n_d^{CNH} + \sum_{d=1}^3 n_d^{CHT}}{\sum_{d=0}^3 n_d^{CNH} + \sum_{d=0}^3 n_d^{CHT}}, \quad (7)$$

$$P_{spec}^2 = \frac{n_0^{CN}}{\sum_{d=0}^3 n_d^{CN}}. \quad (8)$$

Note that P_{spec}^2 equals P^{CN} .

6.1 Results in Static Environment

The BM^S case is discussed in this subsection. T in Subject. 3.2 is set to 50. A 16×16 -sized map is prepared.

Table 4 shows experimental results for the pilot data. They are averaged values of the metrics described first. P^{CHT} is deemed to be an important consideration. The proposed classification achieves as high P^{CHT} as the conventional SOM-based classification [8]–[10] that employs substitution processes il-

Table 4 Averaged results of first described metrics in BM^S case

		Results for the pilot data (%)			
		CN	CNH	CHT	SUS
Actualities	CN	92.11	5.46	2.38	0.04
	CNH	10.63	74.57	14.80	0
	CHT	6.99	15.67	77.22	0.11

illustrated by Figs.5 and 6. The above also apply to P^{CN} .

We have 77.22 percent and 91.46 percent as P_{sens}^1 and P_{spec}^1 on average, respectively. In addition, averaged P_{sens}^2 and averaged P_{spec}^2 equal 91.38 percent and 92.11 percent, respectively. These percentages are also calculated for the pilot data. When generally detecting the hematopoietic tumors, blood physicians comprehensively refer to the results of bone marrow test, lumbar puncture and so forth in addition to the screening data. In other words, it is safely considered that the proposed classification is made, using the barely minimum data (i.e., the screening data). Favorable P_{sens}^1 and P_{spec}^1 are however achieved despite such restricted condition. Assuming that home doctors use the classification system with the aim of roughly recognizing patients to be suspicious, both of sensitivity and specificity (i.e., P_{sens}^2 and P_{spec}^2) exceed 90 percent. In this context, if powerful outcomes associated with lumbar puncture, chest X-ray and so forth are available together with the screening data, it is expected that the probability of missing the tumor patients would be drastically reduced.

An example of the constructed map is shown in Fig.9. A 16×16 -sized map has 15 minimum (i.e., 2×2 -sized) blocks a side. Each of the squares in Fig.9 corresponds to a minimum block, and a label is assigned to it in the manner described in Subsect. 4.2. The above is why the map consists of 225 ($=15 \times 15$) squares. T-BMSOM-based neuron labeling is executed at the block level. The probability that the same label is assigned to each of the neighboring two neurons is therefore high in the BM^S case, compared with conventional SOM-based labeling. Clear clusters of blocks are formed in the map.

Fig.9 Example of map constructed in BMS case

When the class of the presented data is misjudged, the block, which is located in the boundary region between clusters, tends to fire. In other words, if the winner block belongs to the boundary region, medical doctors can make a decision on the class of the presented data, sufficiently keeping in mind the possibility that the true class disagrees with the class specified by the label of the winner. Positions of winner blocks thus serve to guess dependable levels of classification results. To adequately take advantage of this concept, clusters with sharp outlines are preferable to those with ambiguous outlines. T-BMSOM learning yields maps with clear clusters as shown in Fig.9, and hence the proposed classification is useful in visually aiding medical doctors for the tumor diagnosis.

Some of previously proposed algorithms, such as Ward's method, K-means algorithm and c4.5 algorithm, are also applicable to the discussed issue of classifying the screening data. It however seems that they lack the capability of aiding medical doctors in the above visualized way. To the best of our knowledge, the T-BMSOM-based classification is the first algorithm not only fulfilling such visualization requirement but also coping well with environmental changes during learning.

6.2 Results in Dynamic Environments

The BM^{D1} and BM^{D2} cases are explored in this subsection. Parameters are as follows : $T=50$ for training a 16×16 -sized map at each stage, $H=20$ and $M=16$ in the BM^{D1} case, $in=25$ in the BM^{D2} case.

In the BM^{D1} case, subsets are added as shown in Fig.7. First, the order of adding subsets is randomly changed 20 times, and 20 permutations are generated for the data set. T-BMSOM learning and classification are conducted on the condition specified by one of such permutations, and P^{CN} , P^{CNH} and P^{CHT} are calculated at each stage. P_{sens}^1 and P_{spec}^1 , P_{sens}^2 and P_{spec}^2 are also calculated. These are repeated for any of the 20 permutations. Fig.10 depicts averaged values of P^{CN} , P^{CNH} and P^{CHT} at each stage, whereas Table 5 shows averaged values of sensitivity and specificity at the four distinct stages. Note that the numerical results are obtained by presenting pilot data to the map.

P^{CN} , P^{CNH} and P^{CHT} take comparatively low values in

Fig.10 P^{CN} , P^{CNH} and P^{CHT} in BM^{D1} case

Table 5 Sensitivity and specificity at four distinct stages in BM^{D1} case

Stage number	P_{sens}^1 (%)	P_{spec}^1 (%)	P_{sens}^2 (%)	P_{spec}^2 (%)
1	69.03	92.76	85.52	87.07
6	75.05	92.26	90.06	90.36
11	76.20	92.82	91.21	89.40
16	77.48	91.99	91.67	92.30

the first three stages. They however increase as the stage number becomes larger. Finally, T-BMSOM-based classification in the BM^{D1} case achieves as high P^{CN} , P^{CNH} and P^{CHT} as that in the BM^S case. Changes in sensitivity and specificity are similar to the above. Eqs. (5) and (6) (or (7) and (8)) begin to calculate approximately the same values of P_{sens}^1 and P_{spec}^1 (or P_{sens}^2 and P_{spec}^2), at the sixth stage, as they calculate in the BM^S case. It is thus safely considered that the T-BMSOM-based classification is robust to environmental changes caused by sudden increase of the number of the training data. The map is successfully trained without adjustments of time-controlled parameters from the outside, and keeps working well as the classification tool.

Maps constructed in the first, sixth, eleventh, and sixteenth stages are shown in Fig.11 (a)–(d), respectively. Recall that a square in each map corresponds to a 2×2 -sized block. The maps in Fig.11 resemble each other in terms of the connectivity of the formed clusters. This means that the favorable connectivity is well sustained, and T-BMSOM-based incremental learning is adequately executed. Disjoint clusters associated with CNH and CHT appear at the first

(a) 1st stage

(b) 6th stage

(c) 11th stage

(d) 16th stage

Fig.11 Examples of constructed maps in BM^{D1} case

stage. This is due to the fact that attributes of members in the first training data set are insufficient to form three clear clusters corresponding to CN, CNH and CHT. Such disjointed clusters however become smaller as the stage number increases. Note that the number of members in the training data set increases together with the stage number. In Figs.11 (c) and (d), no disjointed clusters are formed. It is thus expected that the advantage of data visualization due to using maps, which has been described in Subsect.6.1, also applies to the BM^{D1} case.

Let us briefly discuss the BM^{D2} case. The following numerical results are obtained by executing a five-fold cross-validation. The averaged P^{CN} , P^{CNH} and P^{CHT} are shown in Fig.12. The high accuracy is achieved for each of these metrics. Recall that the number of items in Fig.12 increases with the stage number. While concrete results are omitted for P_{spec}^1 and P_{sens}^2 , it has been checked that changing trends relative to them in this case are approximately equal to those in the BM^{D1} case.

Fig.13 depicts clusters formed at four distinct stages. Clear clusters appear in the map. While the cluster associated with CNH is split in Fig.13 (c), one of the disjointed clusters consists of just one block. The degree of fission in the BM^{D2} case is lighter than that in the BM^{D1} case. As the above mentioned, the clear clusters are directly linked with the high capability of aiding medical doctors.

T-BMSOM-based classification is thus of use in terms both of numerical performance and data visualization, even if environmental changes successively arise in the number of data or items.

Fig.12 P^{CN} , P^{CNH} and P^{CHT} in BM^{D2} case

Fig.13 Examples of constructed maps in BM^{D2} case

7. Conclusions

This paper has proposed T-BMSOM-based classification of screening data to recognize hematopoietic tumor patients. It constructs a map by T-BMSOM learning, and the data to be classified is presented to the map. T-BMSOM learning adopts the decision-tree-like search and batch process. The former repeatedly chooses a candidate for the winner out of four blocks included in the same candidate having been most recently determined. The latter accumulates degrees and ratios of modifications on neuron reference vectors, and updates all of the vectors once an epoch. They are useful in reducing computational time complexity required for constructing a map. A member of the original screening data set frequently lacks several of the item values. To generate the training data, the proposed method averages non-missing item values of the screening data belonging to each of CN, CNH, and CHT, and the averages substitute for the items without values.

In simulations of dynamic environments, the following cases have been explored: the case where the number of training data changes, and the case where the number of items changes. Numerical results and block clusters in maps, which appear at several stages during learning, have established that the proposed method has the following advantages.

- 1) It achieves the high classification accuracy, and is applicable, even if a map must be trained in dynamic environments.
- 2) It allows classifying the data in the forms of clear block clusters. The maps are also exploited as tools that visually inform inexperienced users of the dependability associated with the classification.

In future studies, the proposed method will be modified to diagnose detailed types of hematopoietic tumors.

References

- [1] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol.43, pp.59-69, 1982.
- [2] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed., Springer-Verlag New York, Inc., New York, 1989.
- [3] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag New York, Inc., Secaucus, NJ, 1995.
- [4] M. Mitihata, T. Miyoshi, and H. Masuyama, "A consideration on the labels of self-organizing map after refresh learning," *Procs. of 6th Int. Conf. on Soft Computing (IIZUKA2000)*, pp.233-238, 2000.
- [5] H. Kurosawa, Y. Maniwa, K. Fujimura, H. Tokutaka, and M. Ohkita, "Construction of checkup system by self-organizing maps," *Procs. of Workshop on Self-Organizing Maps*, pp.144-149, 2003.
- [6] Y. Maniwa, H. Tokutaka, K. Fujimura, M. Ohkita, T. Iokibe, K. Tada "Use of Chaos and Self-Organizing Maps for Acceleration Plethysmogram Information", *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 16, no. 3, pp. 253-261, 2004.
- [7] A. Ohtsuka, N. Kamiura, T. Isokawa, N. Minamide, M. Okamoto, N. Koeda, and N. Matsui, "A self-organizing map approach for detecting confusion between blood samples," *SICE Trans.*, vol.41, no.7, pp.587-595, 2005.
- [8] A. Ohtsuka, H. Tanii, N. Kamiura, T. Isokawa, N. Okamoto, N. Minamide, and N. Matsui, "On Detection of Hematopoietic Tumors Using Self Organizing Maps," *Proceeding of SICE Annual Conference*, pp.2063-2066, Okayama, Japan, Aug. 8-10, 2005.
- [9] N. Kamiura, A. Ohtsuka, H. Tanii, T. Isokawa, and N. Matsui, "On detection of hematopoietic tumors using self organizing maps and genetic algorithms," *Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics*, pp.1925-1930, 2005.
- [10] A. Ohtsuka, H. Tanii, N. Kamiura, T. Isokawa, and N. Matsui, "Self-organizing map based data detection of hematopoietic tumors," *IEICE Trans.*, vol.E90-A, no.6, pp.1170-1179, June 2007.
- [11] A. Ohtsuka, N. Kamiura, T. Isokawa, and N. Matsui, "Self-organizing map based on block learning," *IEICE Trans.*, vol.E88-A, no.11, pp.3151-3160, Nov. 2005.
(2007年6月14日 受付)
(2007年11月12日 採録)

[Contact Address]

2167 Shosha, Himeji, Hyogo, 671-2201, Japan
Graduate School of Engineering, University of Hyogo
Naotake KAMIURA
TEL : 079-267-4918
FAX : 079-267-4918
E-mail : kamiura@eng.u-hyogo.ac.jp

Information about Author

Naotake KAMIURA [non-member]

Naotake Kamiura received his B.E. degree in 1990, M.E. degree in 1992 and Dr. Eng. degree in 1995 from Himeji Institute of Technology. He is currently an associate professor of the Division of Computer Engineering, Graduate School of Engineering, University of Hyogo. His current research interests include design and fault tolerance of multiple-valued logic circuits, fuzzy controllers and neural networks. He is a member of IEEE and Institute of Electronics, Information and Communication Engineers of Japan.

Hirotsugu TANII [non-member]

Hirotsugu Tanii received his B.E. degree in 2005 from Himeji Institute of Technology, Japan and M.E. degree in 2007 from University of Hyogo, Japan. He is currently working for Fujitsu Ten Ltd. His research interests include self-organizing maps and neural computing.

Akitsugu OHTSUKA [non-member]

Akitsugu Ohtsuka received his B.E. degree in 2002 and M.E. degree in 2004 from Himeji Institute of Technology, Japan and Dr. Eng. degree in 2007 from University of Hyogo, Japan. He is currently working for Yamato Scale Co., Ltd. His research interests include self-organizing maps and neural computing.

Teijiro ISOKAWA [non-member]

Teijiro Isokawa received his B.E. degree (Electronic Engineering), M.E. degree (Electronic Engineering), and D.E. degree (Doctor of Engineering) in 1996, 1999, and 2004, respectively, from Himeji Institute of Technology, Japan. He is currently an Associate Professor of the Division of Computer Engineering, Graduate School of Engineering, University of Hyogo, Japan. His research interests include nanocomputing, hypercomplex-valued neural networks, and cognitive models in visual systems. He is a member of the IEEE, Institute of Electronics, Information and Communication Engineers of Japan, and Society of Instrument and Control Engineers.

Nobuyuki MATSUI [non-member]

Nobuyuki Matsui received his B.S. degree in physics from the Faculty of Science, Kyoto University, Japan, in 1975, and M.E. and Dr. Eng. degrees in nuclear engineering from Kyoto University in 1977 and 1980, respectively. He is currently a professor of the Division of Computer Engineering, Graduate School of Engineering at University of Hyogo. He is a member of ISCIE, Institute of Electronics, Information and Communication Engineers of Japan, Society of Instrument and Control Engineers, and the Physical Society of Japan.