2009 Ninth International Conference on Intelligent Systems Design and Applications

Accuracy Improvement of SOM-based Data Classification for Hematopoietic Tumor Patients

Naotake Kamiura, Ayumu Saitoh, Teijiro Isokawa and Nobuyuki Matsui
Dept. of Electrical Engineering and Computer Sciences
Graduate School of Engineering, University of Hyogo
{kamiura, saitoh, isokawa, matsui}@eng.u-hyogo.ac.jp

Abstract

This paper presents map-based data classification for hematopoietic tumor patients. A set of squarely arranged neurons in the map is defined as a block, and previously proposed block-matching-based learning constructs the map used for data classification. This paper incorporates pseudo-learning processes, which employ block reference vectors as quasi-training data, in the above training processes. Pseudo-learning improves the accuracy of classification. Experimental results establish that the percentage of missing the screening data of the tumor patients is very low.

1. Introduction

Blood physicians are specialists in detecting hematopoietic tumors. In regions where the number of blood physicians is limited, home doctors are key persons for the early detection. Blood physicians and home doctors are therefore anxious for breakthrough devices that visually warn users of the possible onset of the tumors.

When learning is successfully complete for a self-organizing map (SOM) [1]-[4], some neuron clusters are clearly formed in the map according to attributes of data. Since processing results are visually provided by neuron clusters, SOM's have recently been used to express medical data [5]-[7]. In [8]-[10], SOM's are adopted to recognize the data of hematopoietic tumor patients. To the best of our knowledge, except for such schemes, no works based on soft computing have been reported to classify blood-test data of normal persons, hematopoietic tumor patients, and non-hematopoietic tumor patients. Especially, block-matching-based SOM's (BMSOM's) [11] are applied in [10]. A set of squarely arranged neurons forms a block, and the winner search is made for the block level. In nonstationary environments, a training data set changes suddenly. Such situations probably occur in clinical practice. Fast BMSOM learning referred to as T-BMSOM[10] constructs maps achieving high accuracy in finding data of the tumor patients, even if maps are in nonstationary environments. Owing to the significance of tumor diagnosis, new mechanism improving the accuracy is required at any time.

This paper proposes a method of improving the classification accuracy for the tumor patients' data, incorporating pseudo-learning in T-BMSOM. A block has a reference vector. Pseudo-learning employs block reference vectors as quasi-training data, whereas regular training data are generated from screening data. Pseudo-learning reduces differences of frequencies of being winners among blocks. Since winner blocks chosen for quasi-training data are known in advance, costs required for winner search are extremely low for pseudo-learning. Thus, while proposed learning requires extra processes related to quasi-training data, the total computational time complexity is within tolerable level. Simulations establish that pseudo-learning is useful in improving the accuracy of precisely judging the data class of tumor patients and that of patients contracting other diseases.

2. Preliminaries

Blood physicians actually use screening data. They are easily available to home doctors. Vectors to be entered into the classification system are therefore generated from them. Each of the data consists of 56 items (e.g., white blood cells and chlorine [8]-[10]). Data classes are as follows: class of normal persons (CN), that of hematopoietic tumor patients (CHT), and that of non-hematopoietic tumor patients (CNH).

A block is a set of neurons arranged in square. One of the blocks is chosen as a winner. In an $N\times N$-sized map, the maximum (or minimum) block size is $(N-1)\times(N-1)$ (or 2×2). A block has a reference vector. It equals the average of reference vectors of neurons in

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.150

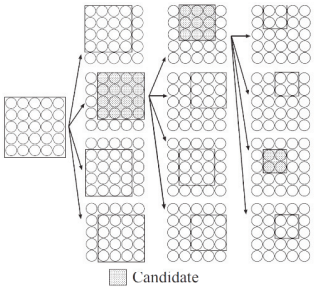

Figure 1. Example of decision-tree-like winner search

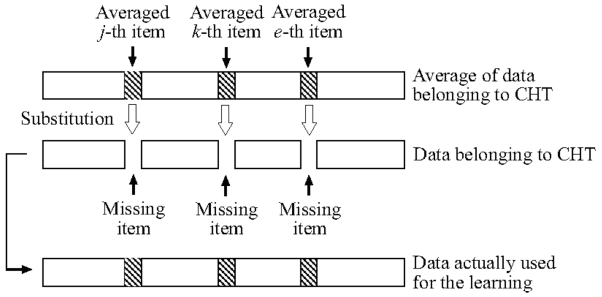

Figure 2. Substitution executed for generation of training data belonging to CHT

the block. To find a winner block, block reference vectors are checked in terms of Euclidean distances to presented input data.

In T-BMSOM learning [10] for an $N\times N$-sized map, the decision-tree-like search chooses $(N-2)$ candidates for winner per member of the training data set. The size of the $(N+1-s)$-th candidate is $(s-1)\times(s-1)$ where $3\le s\le N$. This candidate has the shortest distance to the presented training data among four $(s-1)\times(s-1)$-sized blocks, which are included in the same $s\times s$–sized block chosen as the $(N-s)$-th candidate. The block with the shortest Euclidean distance is finally chosen as the winner for the presented data, out of such $(N-2)$ candidates. Figure 1 depicts an example.

Batch learning is also employed in T-BMSOM. The following formulas are defined for the i-th neuron.

$$W^{i}_{update}(tc,j) \leftarrow W^{i}_{update}(tc,j-1) + x(tc,j)/\alpha, \quad (1)$$

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

where $x(tc, j)$ is the j-th training data on the tc-th epoch, and $1\le tc\le T$. The learning completion condition is specified by T. α is the size of the winner block with the i-th neuron. Every time the training data is presented, the above two values are accumulated. When the tc-th epoch is complete, all neuron reference vectors are updated at once. The i-th neuron reference vector, $W^i(tc)$, is then given by

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

where Q is the total number of the training data.

3. Map-based data classification for hematopoietic tumor patients

3.1. Construction of maps

The screening data generally includes some items without measured values [8]-[10]. Training a map using such incomplete data degrades its classification capability. The imputation is therefore made as follows. A set of the screening data is prepared. Their classes are rightly known. For a set of $(P+U)$ data belonging to some class, let us assume that each of the P data lacks the k-th item value where $1\le k\le 56$. The k-th item values are then averaged for the remaining U data, and the average is given to the k-th item of the above P data. Figure 2 depicts an example. After the imputation is complete for all missing items, the 56-dimensional data is applied as the regular training data.

T-BMSOM learning sometimes fails to update block reference vectors appropriately, if differences of frequencies of being winners are great among blocks. The adaptability of training a map to a nonstationary environment becomes more robust, if the differences can be reduced. Pseudo-learning is one of the promising schemes for reducing the differences. Neuron reference vectors in a $Q\times Q$-sized block are updated, assuming its block reference vector to be the presented training data. Such block reference vector is hereafter referred to as the quasi-training data. The winner for some quasi-training data is clearly the block with it as the reference vector. Thus, since the winner block is known in advance, pseudo-learning does not have to require presenting the quasi-training data actually. The batch process is also applied for pseudo-learning to update neuron reference vectors at once.

Pseudo-learning is conducted before a normal epoch of T-BMSOM starts. An epoch of proposed learning consists of the stage for presenting the quasi-training data hypothetically and that for presenting regular training data. If a 5×5-sized map is prepared and reference vectors of 3×3-sized blocks are employed as the quasi-training data, the number of quasi-training data is 9. For each of quasi-training data, Equations (1) and (2) calculate values related to updating the reference vector of the i-th neuron. The values are accumulated. Once the first stage is complete, all neuron reference vectors are updated, using Equation (3). The second stage in the same epoch follows the first stage, using regular training data.

Let us compare the case of cancelling pseudo-learning with that of adopting it. If nonstationary environmental changes (e.g., update of training data

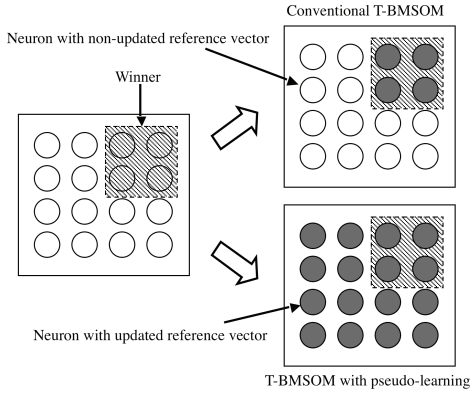

Figure 3. T-BMSOM with pseudo-learning and conventional T-BMSOM

set) occur under executing normal T-BMSOM learning, it is possible that some small-sized blocks are successively chosen as winners. Figure 3 depicts an example. If the upper and rightmost 2×2-sized block is successively chosen as a winner, 12 neuron reference vectors outside it are never updated. Pseudo-learning copes well with such issue. To simplify the discussion, reference vectors of all 3×3-sized blocks are considered to be quasi-training data. The 3×3-sized block shares at least one neuron with the upper and rightmost 2×2-sized block. Therefore, updating neuron reference vectors in this block absolutely affects all 3×3-sized blocks. Pseudo-learning hypothetically presents vectors of the 3×3-sized blocks affected in the above way as quasi-training data, and stores values associated with the modifications on neuron reference vectors in the 3×3-sized blocks. This is why the area of neuron reference vectors to be updated is enlarged in proposed learning, compared with the case where pseudo-learning is cancelled.

3.2. Block labeling and data classification

Once learning is complete, labeling is made for Q×Q-sized blocks as follows. For all items, averages of regular training data belonging to the class r are calculated, where r∈{CN, CNH, CHT}. Let ALD_r denote the data with such averages as item values. Let B_i^Q denote the i-th Q×Q-sized block, and let W_i^Q denote its block reference vector. The three distances, $||W_i^Q - ALD_r||$'s, are calculated for B_i^Q. The label assigned to B_i^Q equals that of the data with the minimum distance to B_i^Q. If the data is equal to other data in distance to B_i^Q, B_i^Q is labeled SUSPENSION ("SUS" for short). For example, when $||W_i^Q - ALD_{CN}||$=5, $||W_i^Q - ALD_{CNH}||$=15 and $||W_i^Q - ALD_{CHT}||$=5, SUS is assigned to B_i^Q as its label.

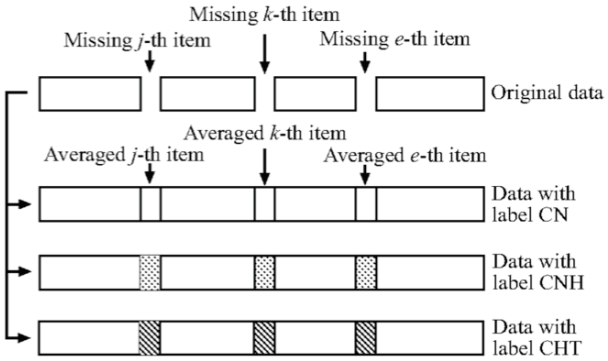

Figure 4. Generation of three data used for classification

Table 1. Example of judgments

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

The screening data unused for learning is referred to as the pilot data. To classify the pilot data, three data are generated from each of the originals. For arbitrary items with no measured values, the averages of the regular training data belonging to the class r are given to the original data as the corresponding item values. Figure 4 illustrates the above. The label assigned to the generated data means the class of the regular training data used to calculate the averages.

The three generated data are presented to a map with labeled blocks. If the presented data has the label different from that of a winner block, a response is disregarded. If the label of one of the three generated data only equals the label of a winner block, only one response is available. In other words, the original pilot data belongs to the class denoted by the latter label. In the case where labels of the two generated data at least equal those of winners, either two or three responses are available. The Euclidean distance is then 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 1 shows an example. It is presumed that the class of the original pilot data is CN. Finally, if there are no available responses, the data class is SUS. This means that the judgment is suspended.

4. Classification model

A set of screening data is divided in H subsets, and any combination of such subsets is employed as a

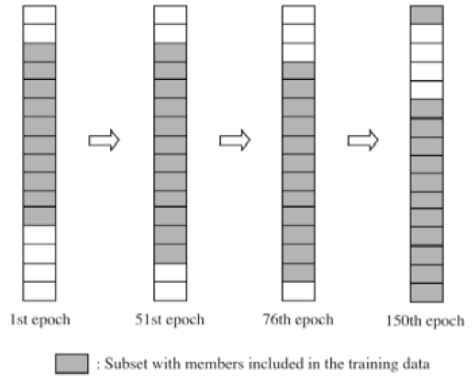

Figure 5. Model of environmental change defined as BM^{NS} case

regular training data set at a given point in time. The number of members in some subset is approximately equal to that in another subset. The imputation in 3.1 is applied to obtain complete data. The following model case is assumed to be a nonstationary environment, and is referred to as the BM^{NS} case.

In the BM^{NS} case, the maximum number of the regular training data is fixed. The regular training data set initially consists of M subsets ($M<H$). It is updated by adding a subset every TD epochs. Let H_{MAX} denote the maximum number of subsets in the data set. After the number of subsets reaches H_{MAX}, a subset is added and one of the M subsets included in the initial data set is deleted. Adding and deleting a subset in this way are successively invoked every TD epochs. The same subset is never simultaneously chosen as the target to be added and that to be deleted. Let ND denote the number of such invocation. Assuming that the above are defined as one trial of the simulation, the total number of epochs is $(H_{MAX}-M+ND+1)TD$ per trial.

Figure 5 depicts an example. We have (H, M, TD, H_{MAX}, ND)=(16, 10, 25, 12, 3). The regular training data set initially consists of the 3rd through 12th subsets at the 1st epoch. A new subset is added at the 26th and 51st epochs. The number of subsets reaches H_{MAX} at the 51st epoch. The 3rd subset is deleted and the 15th subset is added at the 76th epoch. Adding and deleting a subset are made every 25 epochs. The final environmental change occurs at the 126th epoch. It is related to the 6th and 1st subsets. The resultant regular data set is used until the 150th epoch.

The BM^{NS} case corresponds to the situation that old data must be deleted from the storage for any reason, to process new data.

5. Experimental results

The screening data are prepared for 1831 examinees.

Table 2. Actualities versus classification results

		Numbers 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}

The classes of these data are known in advance. For the data of any examinee, about 28 percent of the 56 item values are missing on average. All 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.

Entries in Table 2 denote the numbers of classified data. A column corresponds to a classification result. A row corresponds to an actual class. n_0^{CN}, n_1^{CHN} and n_2^{CHT} therefore mean the numbers of data precisely judged as those belonging to CN, CNH and CHT, respectively. The classification capability is evaluated by the following percentage: the percentage of the number of the data judged as belonging to the class r ($r\in${CN, CNH, CHT, SUS}), compared to total number of the data belonging to the actual class. The percentages associated with precisely judged data are specially denoted by P^{CN}, P^{CNH} and P^{CHT}:

$$P^{CN} = n_0^{CN} / \sum_{d=0}^{3} n_d^{CN}, \quad (4)$$

$$P^{CNH} = n_1^{CNH} / \sum_{d=0}^{3} n_d^{CNH}, \quad (5)$$

$$P^{CHT} = n_2^{CHT} / \sum_{d=0}^{3} n_d^{CHT}. \quad (6)$$

A 16×16-sized map is prepared. Quasi-training data equals all reference vectors of 2×2-sized blocks. When environmental changes occur on a regular training data set, learning is successively conducted without initializing neuron reference vectors. In other words, initial vectors used for learning correspond to the latest neuron reference vectors just before environmental changes occur. For maps trained by the proposed method, 2×2-sized blocks alone are labeled. This is due to the fact that reference vectors of such blocks are quasi-training data. The method for constructing a map in [10] labels blocks with size from 2×2 through 15×15, if the map size is 16×16.

Parameters in the BM^{NS} case are as follows: (H, M, TD, H_{MAX}, ND)=(16, 10, 25, 12, 7). Due to the five-fold cross-validation, five combinations of screening data are available as regular training data sets. A combination consists of 16 subsets, and an initial regular training data set has 10 subsets out of them. Ten sets are generated as initial regular training data sets per combination. The proposed method is evaluated, considering one of the initial sets to be a starting point. The pilot data are classified every 50 epochs of training a map. Once an evaluation is

Table 3. Averaged results achieved by proposed method in BM^{NS} case

(a) 50 epochs later

		Results for the pilot data(%)			
		CN	CNH	CHT	SUS
Actualities	CN	96.17	0.68	3.15	0.00
	CNH	6.85	79.61	13.54	0.00
	CHT	5.08	13.92	81.00	0.00

(b) 100 epochs later

		Results for the pilot data(%)			
		CN	CNH	CHT	SUS
Actualities	CN	95.98	0.84	3.17	0.02
	CNH	6.76	79.59	13.65	0.00
	CHT	4.85	14.28	80.87	0.00

(c) 150 epochs later

		Results for the pilot data(%)			
		CN	CNH	CHT	SUS
Actualities	CN	96.12	0.87	2.99	0.02
	CNH	6.13	79.91	13.95	0.00
	CHT	4.95	14.52	80.53	0.00

(d) 200 epochs later

		Results for the pilot data(%)			
		CN	CNH	CHT	SUS
Actualities	CN	95.80	0.80	3.38	0.02
	CNH	6.29	80.70	13.01	0.00
	CHT	5.02	14.45	80.53	0.01

(e) 250 epochs later

		Results for the pilot data(%)			
		CN	CNH	CHT	SUS
Actualities	CN	96.45	0.75	2.78	0.02
	CNH	6.33	80.47	13.21	0.00
	CHT	5.19	13.81	81.00	0.00

complete, the above are repeated, using one of the remaining training data sets. Table 3 shows averages obtained from all evaluations for pilot data.

Let us specially discuss the classification results of CHT-class data. The classification capability of such data is demonstrated with the bottom rows in Table 3. When generally detecting the tumors, blood physicians comprehensively refer to powerful outcomes (e.g., bone marrow test and lumbar puncture) in addition to the screening data. The screening data are thus the barely minimum data. While the classification is made under such restricted condition, P^{CHT}'s exceed 80.5 percent. If it is acceptable that CNH and SUS are pseudopositive, the probability that the proposed method fails to notice the CHT-class data is less than 5.2 percent. The dangerous false classification must be avoided with great care for the CHT-class data. In the above context, if the powerful outcomes are available together with the screening data, the probability of missing the tumor patients is drastically reduced.

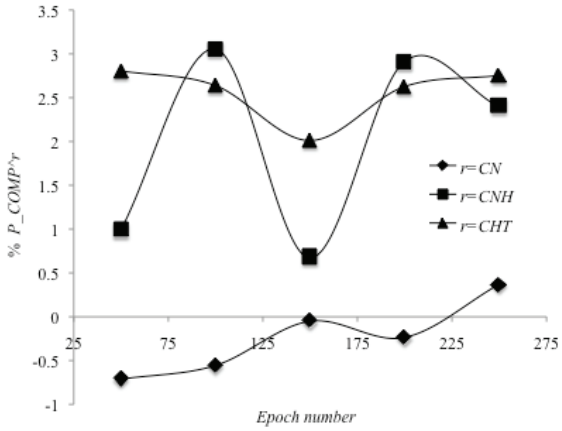

Figure 6. Comparison of proposed method with other classification in [10]

Let us next compare the proposed method with the method in [10], provided that parameters on environmental changes and way of data classification are similar to those in the above BM^{NS} case. The following percentage is assessed: $P_{COMP}{}^{r}=(N_{PRM}{}^{r}-N_{OTM}{}^{r})/N_{OTM}{}^{r}$, where $N_{PRM}{}^{r}$ (or $N_{OTM}{}^{r}$) denotes the numbers of data whose class is precisely judged as r ($r\in${CN, CNH, CHT}) by the proposed method (or method in [10]). Figure 6 depicts averaged results.

The proposed method is almost equivalent to the method in [10] in terms of finding the CN-class data. The former however surpasses the latter in recognizing the other class data. The former requires about 20 percent more training data than the latter per epoch in the above BM^{NS} case, owing to 225 quasi-training data. However, since winners for quasi-training data are clearly known in advance, the cost of presenting one of them is much less than that of presenting one of the regular training data. In addition, the proposed method has the advantage of low cost of block labeling. This is because the proposed method presents 675 data to finish labeling, whereas the method in [10] presents 3717 data in the above BM^{NS} case. The proposed method thus achieves performance advances in Figure 6 in return for the slight increase of total costs.

Let us finally discuss block clusters. Figure 7 depicts an example of the trained map in the above BM^{NS} case. A map has 15 2×2-sized blocks a side. The small-sized squares correspond to such blocks. Positions of the winners serve to guess the dependable levels of judgments made by the map [8]-[10]. If the winner block belongs to the boundary region between clusters, medical doctors can make a decision for the presented data, sufficiently keeping in mind the possibility that the true class disagrees with the class

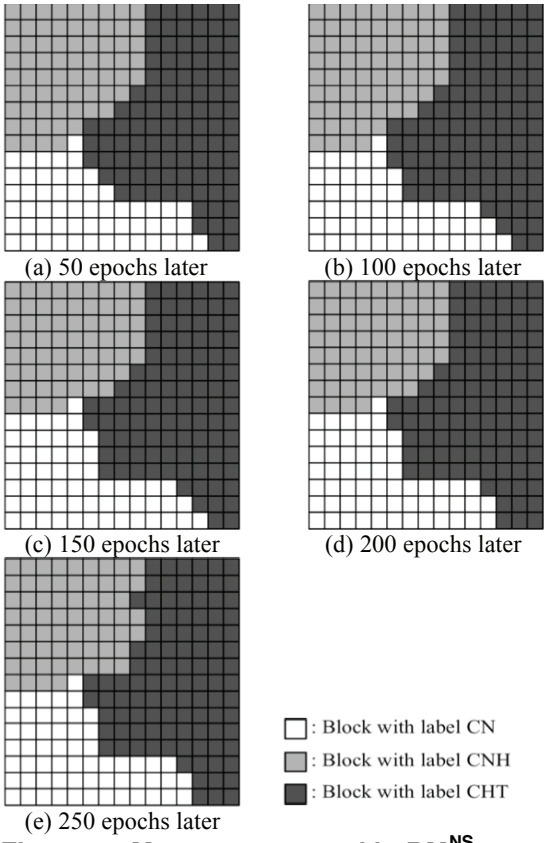

Figure 7. Maps constructed in BM^{NS} case

specified by the label of the winner. To adequately take advantage of this concept, clusters with sharp outlines should be formed. The proposed method yields maps with clear clusters as shown in Figure 7. Thus, pseudo-learning never spoils the capabilities of maps in visually aiding medical doctors.

6. Conclusions

This paper has proposed a map-based method for recognizing screening data of hematopoietic tumor patients with high accuracy. Pseudo-learning is adopted to improve the plasticity of trained maps to nonstationary environments. It employs block reference vectors as quasi-training data, and presents them hypothetically. Since the cost of presenting such data is much less than that of presenting the regular training data, the extra computational time complexity caused by pseudo-learning is tolerable. Experimental results in a nonstationary environment establish that pseudo-leaning makes it possible to improve the probability of judging the CNH-class and the CHT-class data.

In future studies, the proposed method will be modified to diagnose detailed types of the 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] 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.
[9] 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.
[10] N. Kamiura, H. Tanii, A. Ohtsuka, T. Isokawa, and N. Matsui, "Classification for data of hematopoietic tumor patients with fast block-matching-based self-organizing map learning in dynamic environments," Journal of Japan Society for Fuzzy and Intelligent Informatics, vol.20, no.1, pp.66-78, 2008.
[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.