

Visual Stability Improvement of SOM's Feature Map by Initial Value Assignment

MOMOI Shinji

Department of Media Informatics,
Graduate School of Science and Technology,
Ryukoku University
t11m090@mail.ryukoku.ac.jp

MIYOSHI Tsutomu

Department of Media Informatics,
Faculty of Science and Technology,
Ryukoku University
mijosxi@i.ryukoku.ac.jp

Abstract— In SOM learning, learning result depends on initial value of feature map and the location of the node or the distance between nodes on feature map is important factor to determine feature of individual data. For example, in data detection of hematopoietic tumors, given data is detected by location of the feature map. In this paper, we focused on visual stability of SOM feature map, and we proposed new initialization method of SOM feature map. The purposes of proposed method are improvement of visual stability of SOM feature map, and utilization of generalization ability of SOM. By experiments, proposed method is visually stable than conventional method in the point of feature map location, and the computational complexity of proposed method is greatly reduced.

Keywords—component; self-organizing map; feature maps; visual stability; improvement method;

I. INTRODUCTION

Kohonen's Self Organizing Map (SOM) involves neural networks, for which an algorithm learns the feature of input data through unsupervised, competitive neighborhood learning. The SOM is applied in many fields and has been widely studied [4-7]. Based on the conventional SOM learning algorithm, SOM learning is influenced by the sequence of learning data and the initial feature map. The location of the node or the distance between nodes on feature map is important factor to determine feature of individual data.

In conventional method, initial value of feature map has set at random, so a different mapping appears even by same input data. Consequently, it can be said that it lacks visual stability. For example, in data detection of hematopoietic tumors, given data is detected by location of the feature map. If given data are near to the area of physically unimpaired person then the data consider to normal, and if the data are near to the area of tumors then the data consider to abnormal. As well as this example, in many applications the judgment is made by a person depending on a location on the feature map.

If the learning results are visually unstable, maps must become different impressions in each learning, and maps of different impressions could be increased to the same data in different diagnosis.

To solve this visually unstable problem, there are steadies, that euclidean distance of learning data is pre-calculated and vector of feature map permute based on the result of pre-calculate. Because it is necessary to compare distance of all vectors, however, the computational complexity becomes very high, and a lot of time will be spent in analysis of learning data. As a result, generalization ability of SOM is not fully utilized. Moreover, the method focused on high speed leaning, so, there are not enough discussion about visual stability.

In this paper, we focused on visual stability of SOM feature map, and we proposed new initialization method.

II. SELF-ORGANIZING MAP

Kohonen's self-organizing map (SOM) [1] involves of neural networks, that learn the features of input data through unsupervised, competitive neighborhood learning. The SOM is mapping from high to low dimensional space, usually as a two-dimensional (2D) map. It provides a feature map that arranges similar classes nearer to one another to visualize high-dimensional information in a 2D feature map. Map representation makes it easier to understand relations between data. (Fig.1)

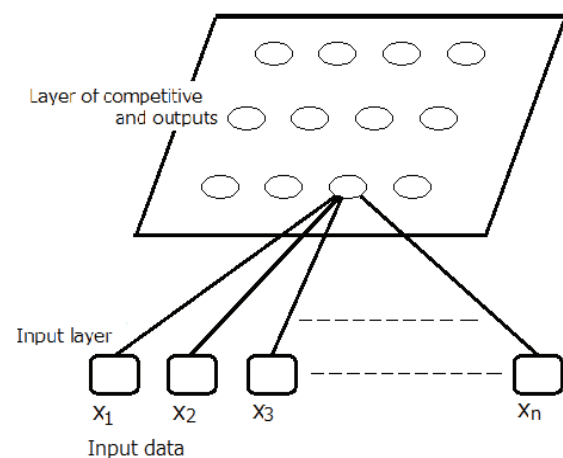


Figure 1. SOM overview

The SOM generally has two layers, an input layer and an output layer. Output layer nodes usually form a 2D grid and input layer nodes are fully connected with those at output layer nodes. Each connection has a connection weight, so all output layer nodes have patterns or vectors to be learned.

After learning, each node represents a group of individuals with similar features, the individual corresponding to the same node or to neighboring nodes. That is, the SOM configures output nodes into a topological representation of original data through a process called self-organization.

In learning process, when an input pattern or input vector is presented to the input layer as learning data, output layer nodes compete mutually for the right to be declared the winner. The winning node is the output layer node whose incoming connection weights are the closest to the input pattern in Euclidean distance. The connection weights of the winning node and its neighbor nodes are then adjusted, i.e., moved closer toward the input pattern.

As learning process progresses, the learning rate and the size of the neighbor area around the winning node decreases, so in an early stage of learning process, large numbers of output layer nodes are adjusted strongly and, in the final stage, the winning node alone is adjusted weakly.

III. PROBLEM OF CONVENTIONAL TECHNOLOGY

A. Initial Value of Reference Vector

Vector of SOM's feature maps (reference vector) has been initializing by random number on conventional technology as described ahead. However, in learning process of SOM, simply initializing by random number has a big influence chiefly first learning. The example of possible problem in case of simply initializing it by random number is enumerated as follows.

First of all, there are large change of coordinates of winning node even if used same leaning data, because value of reference vector is selected by random number. In different learning, the location of winning node for the same leaning data is usually different, because initial vector is randomly set its values. The example is shown in figure 2.

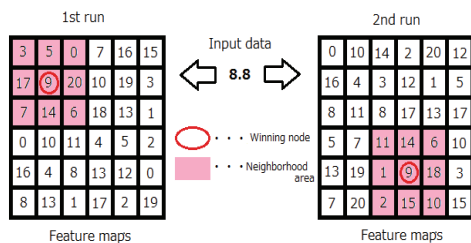


Figure 2. Difference of coordinates of winning node by random number

In the example, feature maps is 6×6, input data is 8.8, and neighborhood area is one surrounding space. Whenever winning node becomes completely random when value of reference vector is completely random like this. Vector surrounding winning node, i.e., vector in neighborhood area, also have random value, so, influence of learning is different in each leaning.

B. Data Pre-calculation Method

There are a method which initial value is calculated from input data. The method proposed by Mu-Chun Su [2] is enumerated as an example. (1) four vectors whose distance is the longest in input data are extracted. (2) it is allocated in the four corners of feature maps as weight. (3) data that is the furthest from both ends is allocated to opposite weight. (4) until all weights are assigned, this can be done clockwise repeatedly. As for the method of analyzing input data like this example and deciding initial value, the following problems are enumerated.

First of all, by this initialization method, learning data are analyzed their feature, so SOM leaning is used to miner adjustment. The purpose to use SOM is to analyze what kind of relations are exist in input data. The purpose might be lost by separately analyzing input data to decide initial value of reference vector.

Second, there is a problem that the computational complexity depends on the number of input data and number of dimensions. Therefore, a lot of procedures will be added before it learns by SOM.

Third, the method focused on high speed leaning, so, there are not enough discussion about visual stability.

C. Node Exchange Method

In the research of Miyoshi[3], they proposed initialization method that tried to utilize generalization ability of SOM. The method is to purpose at the speed-up study by improving the procedure of deciding initial value at random. This method relates input vector space to the position of feature maps in exchanging nodes of feature maps referring to learning data at initialization. The average moved distance of all nodes is shortened when this method is used, and the speed-up of the learning speed is confirmed.

In point of view that utilize generalization ability of SOM, this method has same purpose, however, the purpose to improve visual stability of output result is not discussed enough in the research.

IV. PROPOSED METHOD

The purposes of proposed method are (1) improvement of visual stability for output result, and (2) utilization of generalization ability of SOM. To achieve second purpose, range of vector values are calculated from small number of learning data. And to achieve first pupose, reference vectors are sorted and allocated to give feature maps location tendency.

A. Range of Values

To reduce the complexity of learning, we try to limit the range of random values. To determine the range of values, randomly selected from a small number of learning data and approximate the maximum and minimum values for these data. Number of randomly selected data is determined by following formula:

$$n = \frac{N}{\left(\frac{E}{Z}\right)^2 \left(\frac{N-1}{P(1-P)}\right) + 1} \quad (1)$$

where, n is number of randomly selected data, N is a population, E is maximum error, Z is reliability coefficient, and P is population proportion. Using this formula, number of randomly selected data can be kept at a constant value.

Then the random numbers are provisionally allocated as initial value of reference vector. The advantage of this method is shown below. By this algorithm, it is possible to make the range of reference vector correspond to the range of learning data, and the calculation cost doesn't hang so much.

B. Vector Sorting

To give feature map location tendency, reference vector are sorted and allocated to feature map. First, the average of all dimensions value of each reference vector are calculated and, by using these averages as sorting key, reference vectors are sorted. Averages is calculated by the following formula:

$$A = (m_{ij0} + m_{ij1} + m_{ij2} + \dots + m_{ijk}) / k \quad (2)$$

where, A is average, m is reference vector, i is x axis of feature map, j is y axis of feature map, and k is value of dimension.

Then, vectors are allocated to feature map in sorting order. The image of sorting is shown in figure 3.

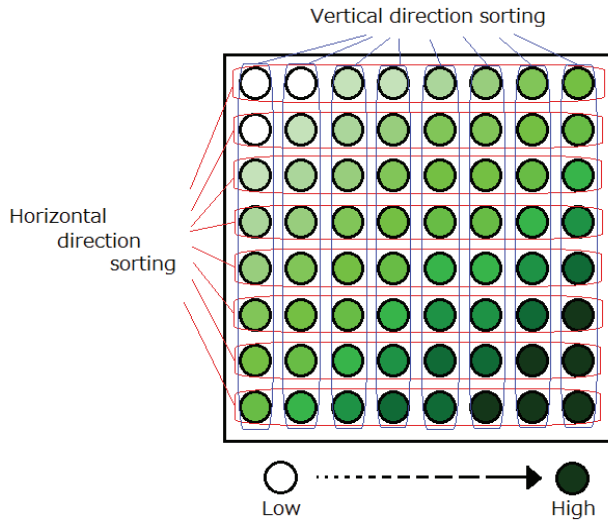


Figure 3. The image of sorting

As a result, the difference of winning node to same learning data is reduced. So, it is thought that output result is become steady.

C. Algorithm of Feature Map Initialization

Algorithm of Proposed Method is following.

- From the learning data, randomly select a value.
- From the selected value to determine the minimum and maximum values.

- random numbers are generated at the range of minimum and maximum values.
- Generated random number are temporarily assigned to the reference vector.
- To calculate the average value of the reference vectors.
- Based on the average, reference vector is done to quick sort in ascending order for the vertical direction feature map.
- Based on the average, reference vector is done to quick sort in ascending order for the horizontal direction feature map.
- Feature maps are made by this sorting, is initial value.

V. EXPERIMENTS FOR VISUAL STABILITY

To compare visual stability between conventional method and proposed method, we performed experiments. Conventional method is that of having initial value of reference vector by random numbers. The experiments are giving the same learning data to these two programs. In order to simplify the visual changes, bi-polarization data was used. The image of learning data is shown in figure 4.

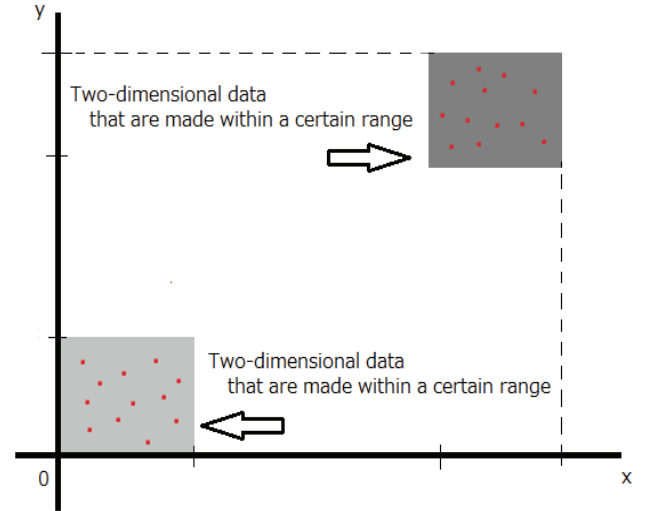


Figure 4. The image of learning data

Experiment parameters are followings. Learning data set is 3*200, feature maps size is 30*30 nodes, learning repetition is 4000 times, neighborhood area is 20 (it decreased from 20 to 1), and learning rate is 0.01*(neighborhood area size).

A. Learning Results

The experimental results is shown in figure 5. The left hand side are results of conventional method and, the right hand side are results of proposed method.

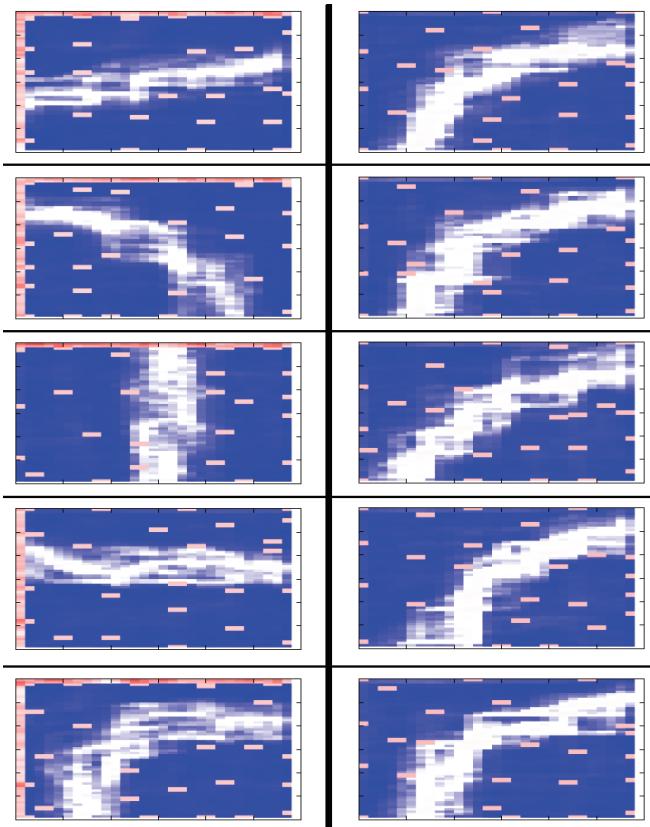


Figure 5. Output result of existing technology and proposal technique

According to the results of conventional method, visually different feature maps are generated at every single learning, even if leaning data is completely same. So, winning node of same input data is located in different part of each feature map. On the other hand, according to the results of proposed method, visually similar feature map are generated by same learning data. So, winning nodes of same input data are located in similar part of feature maps. The position of all input data showed similar tendency, typical winning nodes location are shows in figure 6.

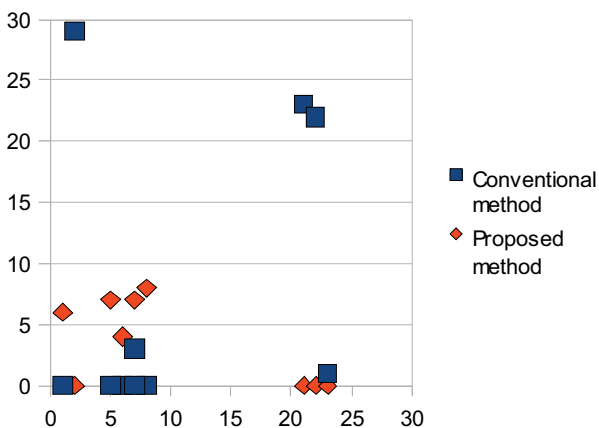


Figure 6. Feature map coordinates on a single input data

Figure 6 shows the location of winning nodes in five different feature maps for a single input data. X axis is x-coordinate and y axis is y-coordinate of feature map. According to Figure 6, winning node locations of conventional method are widely scattered even though input data is same. On the other hand, winning node locations of proposed method is almost unchanged corresponding to same input data.

The computational complexity of conventional method is estimated to $O(n!)$, and the one of proposed method is estimated to $O(n \log n)$. So, it can be said that the computational complexity of proposed method is greatly reduced.

VI. CONCLUSION

In this paper, we proposed new initialization method of SOM feature map. The purposes of proposed method are improvement of visual stability for output result, and utilization of generalization ability of SOM. The range of reference vector values are calculated from maximum and minimum values approximated by randomly selected small number of learning data, and reference vectors are sorted and allocated to give location tendency in feature map.

By experiments, followings are revealed that, proposed method is visually stable than conventional method in the point of feature map location, the computational complexity of proposed method is greatly reduced.

As for further research, it is necessary to think about sorting method not related to number of dimensions. The average is sorting key in this time. However, sorting key to which stability of output result is kept even if the dimension increases is preferable. In addition, this study used a graph to show minutely the stability of the output information on input data. However, we need to think interface to output result more visually clear.

REFERENCES

- [1] Teuvo Kohonen, Self-Organizing Maps, Springer Verlag, 1995.
- [2] Mu-Chun Su, Hsiao-Te Chang, "Fast Self-Organizing Feature Map Algorithm" IEEE Transactions on Neural Networks, Vol.11, No.3, 2000, pp.721-733.
- [3] Tsutomu Miyoshi, "Relation Organization of SOM Initial Map by Improved Node Exchange," Journal of Computers (JCP), ISSN 1976-203X, Vol.3, No.9, 2008, pp.77-84.
- [4] SHIEH Shu-Ling, LIAO I-En, "A New Clustering Validity Index for Cluster Analysis Based on a Two-Level SOM" IEICE transactions on information and systems, Vol.92, No.9, 2009, pp.1668-1674.
- [5] SHIEH Shu-Ling, LIAO I-En, HWANG Kuo-Feng, CHEN Heng-Yu, "An Efficient Initialization Scheme for SOM Algorithm Based on Reference Point and Filters" IEICE transactions on information and systems, Vol.92, No.3, 2009, pp.422-432.
- [6] KATO Satoru, HORIUCHI Tadashi, ITOH Yshio, "A Study on Clustering Method by Self-Organizing Map and Information Criteria" Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol.21, No.4, 2009, pp.452-460.
- [7] Jun YoungPyo, Yoon Hyunsoo, Cho JungWan, "L* Learning: A Fast Self-Organizing Feature Map Learning Algorithm Based on Incremental Ordering" IEICE transactions on information and systems, Vol.76, No.6, 1993, pp.698-706.