Improvement of visual stability by adjusting initial map of SOM and using hierarchical clustering

MIYOSHI Tsutomu[1] MOMOI Shinji[1]

[1] Department of Media Informatics, Ryukoku University, Shiga, JAPAN

Abstract

SOM learning is influenced by the sequence of learning data and the initial feature map. In conventional method, initial value of feature map has set at random, so a different mapping appears even by same input data, so different impressions could be increased to the same data in different diagnosis. In this paper, we focused on visual stability of SOM feature map, and we proposed new initialization methods of SOM feature map and new map clustering method using hierarchical clustering. By experiments, proposed methods are visually stable than conventional method in the view of node mapping.

Keywords: self-organizing map, feature maps, visual stability, improvement method, clustering.

1. Introduction

The SOM is applied in many fields and has been widely studied. 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 give 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 [2], 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, as a result, generalization ability of SOM is not fully utilized.

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

2. Self-Organizing Map

Kohonen's self-organizing map (SOM) [1] involves of neural networks that learn

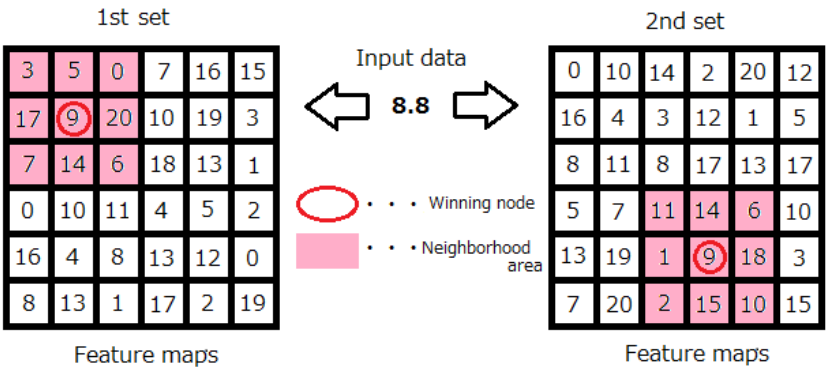

Fig1: Location of winning node in different initialization.

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.

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.

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.

3. Problem of Conventional Method

3.1. Initial Value of Reference Vector

Reference vector of SOM's feature map have been initialized by random number on conventional method as described ahead. However, random initialization has a big influence chiefly beginning of learning. There are big changes in locations of winning node even if used same learning data. In different learning, the location of winning node for the same leaning data is different, because initial vector is randomly set its values. Vector surrounding winning node, i.e., vector in neighborhood area, also have random value, so, influence of learning is different in each learning. The example is shown in figure 1.

3.2. Data Pre-Calculation Method

There is 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) this can

be done clockwise repeatedly until all weights are assigned. As for the method of analyzing input data like this example and deciding initial value, the following problems are enumerated.

First, by this kind of initialization method, learning data are analyzed their feature, so SOM leaning is used to miner adjustment. The purpose to use SOM is to know what kinds of relations exist in input data by using neural network ability. 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 applying SOM learning.

Third, this kind of method mainly focused on high speed leaning, so, there is not enough discussion about visual stability.

4. Proposed Method

The purposes of proposed method are (1) improvement of visual stability of learning result, and (2) utilization of generalization ability of SOM. To achieve these purposes, reference vectors are sorted and allocated to give feature maps location tendency.

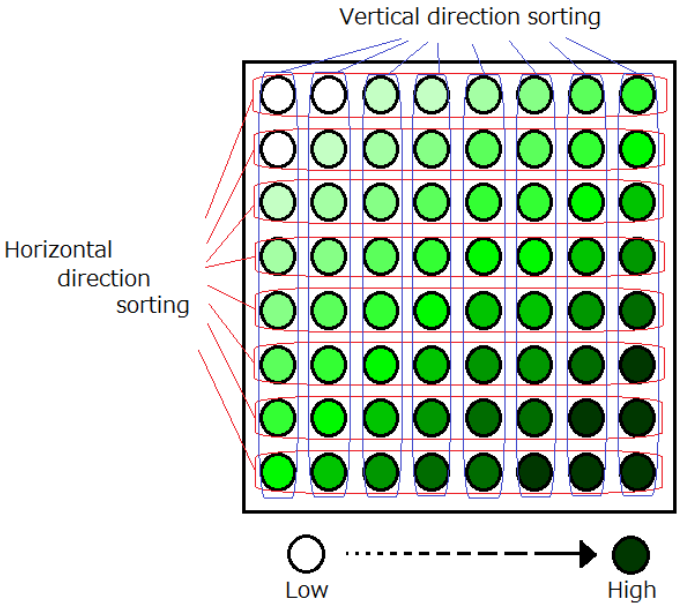

Fig.2: The image of location tendency

4.1. 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. If there are the vectors that average of each dimension value are the same, we compared value lexicographically, i.e., from the first dimension value of the vectors to the end, and the vectors are sorted in ascending order.

Then, vectors are allocated to feature map in the order. The image of location tendency is shown in figure 2.

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

At the learning process, learning data are also extracted in lexicographically order.

4.2. Automatic Clustering

For less clustered data, that is difficult to classify or class bounds are ambiguous, SOM feature map becomes visually ambiguous. To help clear visual classify, we propose to apply clustering method to learned SOM map.

First, we cluster the learning results of SOM by hierarchical clustering method "Ward method". Next, we determine the number of clusters automatically by the amount of maximum increase in the within-groups sum of squares.

5. Experiments

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.

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).

5.1. Node Mapping

The experimental results are shown in figure 3. The left hand side is results of conventional method and the right hand side is results of proposed method.

According to the results of conventional method, visually different feature maps are generated at every single learning even if learning data is completely same. On the other hand, according to the results of proposed method, visually similar feature map are generated by same learning data. So, winning node of same input data is located in similar part of feature maps.

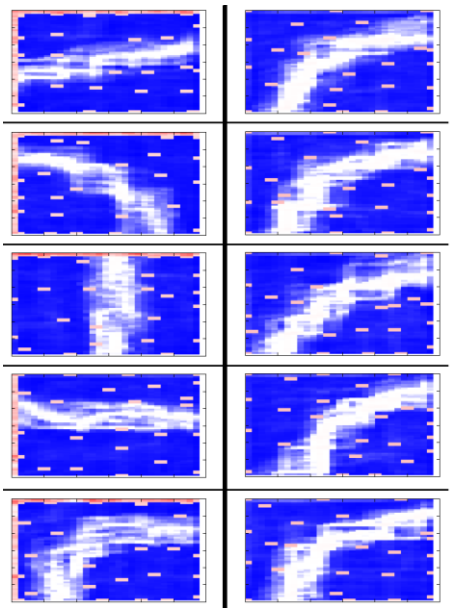

Fig.3: Learning results of conventional method and proposal method.

5.2. Stability Estimation Using Index Data

By above mentioned experiments, we revealed that proposed method is visually stable than conventional method. In this chapter, we revealed stability estimation by winning node location.

This experiment uses index data selected from 3*200 learning data that used in above experiments. Selected index data are (48, 46, 48). The procedure of this experiment is shown below. While 4000 leaning, 1) if the index data is selected as learning data, we keep record of its selection frequency and winning node location in feature map. 2) Then, most frequently

selected winning node location is detected from them.

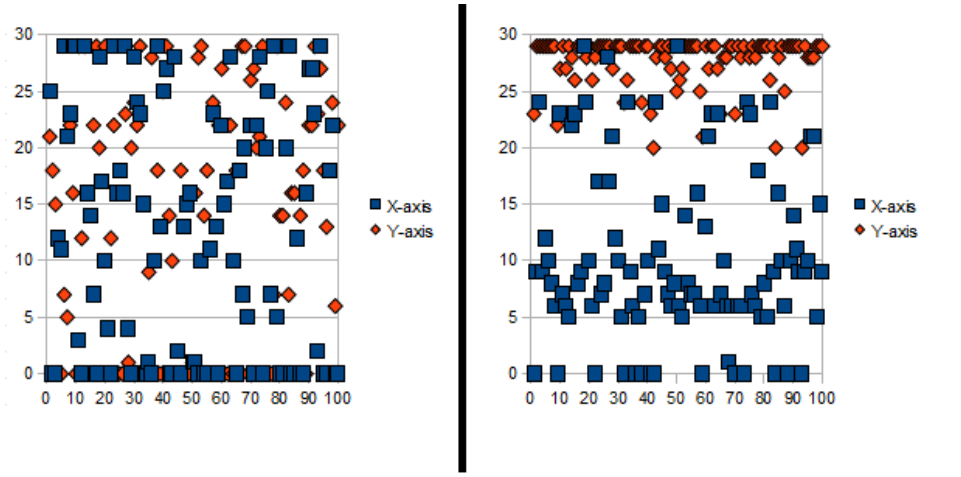

Fig.4: Coordinates of index data in each run.

The experimental result is shown in figure 4. This experiment was done 100 times, and depicted the respective X and Y coordinates. The vertical axis is position, blue square points are coordinates of X, red diamond shape points are coordinates of Y, and the horizontal axis shows the number of experiments. If the points are gathering for each color, coordinate or position have the stable trend. However, if the points are distributing for each color, coordinate or position have the unstable trend.

The left graph of figure 4 shows the conventional methods, the right graph shows the proposed method. According to figure 4, in proposed method, winning node locations are similar, but in conventional method, winning node locations are distributing in each run.

5.3. Clustering

Using 100 SOM learning results, that are difficult to classify visually, determine the number of clusters by proposed automatic clustering.

The experimental results are always divided to mainly 4 clusters and occasionally 3 or 5 clusters automatically. As a result, clustering of proposed method can support of visual understanding.

6. Conclusions

In this paper, we proposed new initialization method of SOM feature map. The purposes of proposed method are improvement of visual stability of learning result, and utilization of generalization ability of SOM. Reference vectors are lexicographically sorted and allocated to give location tendency in feature map. In addition, we proposed a method to clustering the learning results of SOM by "Ward method". By the experiment, followings are revealed that, proposed method is visually stable than conventional method in the view of node mapping.

7. References

[1] Teuvo Kohonen, Self-Organizing Maps, Springer Verlag, 1995.
[2] Mu-Chun Su, Hsiao-Te Chang, “Fast Self-Organizing Feature Map Algo-

rithm" IEEE Transactions on Neural Networks, Vol.11, No.3, 2000, pp.721-733.