Non-stop Learning: A New Scheme for Continuous Learning and Recognition

Yuki Kamiya*, Shen Furao* and Osamu Hasegawa†‡
*Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama, Japan
†Imaging Science and Engineering Lab., Tokyo Institute of Technology, Yokohama, Japan
Telephone: +81-45-924-5180, Fax: +81-45-924-5175, E-mail:hasegawa@isl.titech.ac.jp
‡PRESTO, JST

***Abstract*— Continuous learning and recognition in a fluctuating environment are difficult tasks for traditional pattern recognition techniques. This condition causes the well-known "symbol grounding problem", which is considered to be a salient problem in the field of machine learning that must be solved. This paper presents a new scheme for continuous and incremental learning and recognition based on growing and competitive neural networks. The scheme consists of layered multiple neural networks for data analysis and concept acquisition. In this scheme, the learning phase and recognition phase are not separated. Details of the scheme are shown in this paper along with some results of simulation.**

I. Introduction

In designing artificial intelligence systems that are useful in the real world, we are confronted by a crucial obstacle – the "symbol grounding problem". This problem is caused by the absence of an effective interface between the real world, as an open system, and the systems of reasoning and determination of a plan of action, as symbolic systems.

In a fluctuating environment, such as that in the real world, many difficulties hinder traditional pattern recognition techniques. They can prevent learning and recognition of audiovisual pattern information. On a signal level, a system must cope with pattern signal degradation, the influence of noise, and other hindrances.

On higher-level thinking, although a learner generally learns based on a limited learning sample that is established a priori, if the environment is changed and one recognizes pattern data which differ vastly from learning samples, the recognition result is uncertain. For this reason, it is very difficult to approximate true data distributions using only limited learning samples in such environments. Furthermore, the learner must learn classes incrementally.

To solve the above-mentioned problems, two almost identical algorithms – Growing Neural Gas (GNG) [1] and the Dynamic Cell Structures [2] – were proposed. They operate in non-stationary data distributions and on-line learning. They combine the concepts of vector quantization and continuous neighborhood adaptation. Network nodes compete for determining the node with the highest similarity to the input pattern. A counter link with the winner is increased by the error of the network. According to the time average, nodes with high errors serve as a criterion to insert a new node. The major drawbacks of these methods are their permanent increases in the number of nodes and the drift of the centers to capture the input probability density [3]. Thresholds such as the maximal number of nodes predetermined by use as well as the insertion criterion depend on the overall error or on a quantization error. They are not appropriate because appropriate figures for these criteria cannot be known a priori.

The above-mentioned methods are not suitable for the much harder problems of non-stationary data distributions. The fundamental issue for such problems is: How can a learning system adapt to new information without corrupting or forgetting previously learned information? – the so called stability-plasticity dilemma [4]. Using a utility-based removal criterion, GNG-U deletes nodes that are located in regions with a low input probability density [5]. This criterion serves to follow a non-stationary input distribution, but the former learned prototype patterns are thereby destroyed.

Lim and Harrison proposed a hybrid network that combines advantages of Fuzzy ARTMAP and the Probabilistic Neural Networks for incremental learning [6]. It achieved significantly better results than a regular Fuzzy ARTMAP on Gaussian source separation and on a noisy waveform recognition task. Hamker proposed an extension to the GNG that allows it to learn its number of nodes needed to solve a current task and to dynamically adapt the learning rate of each node separately. Both methods can work for some supervised on-line learning or life-long learning tasks. Nevertheless, they and other supervised learning methods are unsuitable for the constant learning task, which is the topic for this study; it cannot append a label to all signals.

Shen proposed Adaptive Neural Gas (ANG), an extension to the GNG that enables it to fit a non-stationary data distribution [7]. It achieves this task through a combination of a similarity threshold and a local accumulated error. The usage of a utility parameter – error-radius – allows both the judgment of whether the insertion is successful and the control of increasing of nodes. Moreover, using an on-line criterion for node removal, the data set is classified well and noises are eliminated periodically. Although this algorithm is able to address some difficulties of on-line or life-long unsupervised learning, it cannot separate some data distributions with a high-density overlap [8].

This study is intended to design a constant learning and recognition system in fluctuating environments such as the

real world. The objective is to acquire concepts, operating continuously, on-line, and in the real world. The novelty of the system is in its architecture, which combines supervised learning and unsupervised concept acquisition.

This paper describes details of the proposed algorithm. The system is intended to produce a means to acquire concepts independently, epitomizing results of unsupervised classification. It engenders efficient segmentation of distributions with high-density overlap using the limited labeled signals.

II. Proposed Algorithm

A. Overview of the proposed method

For the constant on-line learning task, we intend to produce the proposed algorithm construct with a conceptual structure from the input unlabeled non-stationary data using no prior knowledge such as how many classes exist. We further intend to approximate the conceptual structure that the teacher has, using limited input-labeled data.

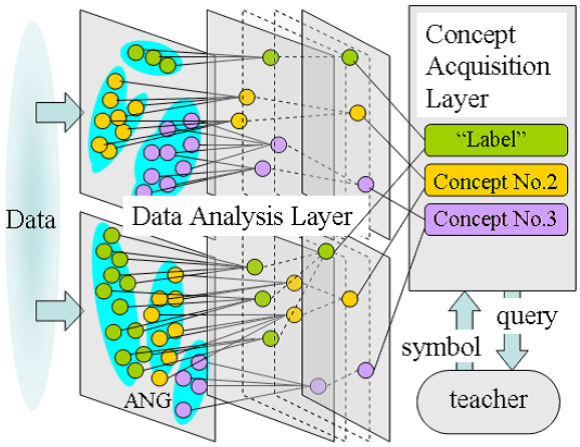

Fig. 1. Overview of the propose algorithm

A schematic overview of the proposed method is shown in Fig. 1. This method consists of layered multiple neural networks for data analysis and concept acquisition.

The data analysis layer has two or more layers and learns data distributions of the input unlabeled non-stationary signals. The bottom layer of this layer is a layer with the fundamental feature of the ANG. Therefore, this layer operates on-line unsupervised classification, learns topology, and eliminates noise for the input signals. Other layers within this layer epitomize the result of the bottom layer. They abstract the topology of the input signals using some prototypes for each class. The upper layer abstracts with fewer prototypes. Moreover, two or more data analysis layers are arranged in parallel to analyze the multiple features cooperatively.

The concept acquisition layer forms concepts of the input unlabeled signals from results of the multiple data analysis layers. By acquiring input labeled signals from a human teacher, either actively or passively, it relates a symbol with each of the created concepts to recognize the following input signals. In addition, it corrects the structures in the data analysis layers and the weight of connections between the data analysis layers and the concept acquisition layer, based on limited input labeled signals.

The above-mentioned system performs bottom-up, on-line, and autonomous concept acquisition along with periodic and top-down supervised learning. The result of that learning is utilized directly for recognition of the following input signals. In this manner, this system combines the learning process and the recognition process. In other words, this system can cope with some difficulties of constant on-line learning, such as high-density overlaps, very numerous input signals, limited input labeled signals, and noise.

B. Bottom-up concept acquisition

We shall explain the way in which a result of on-line unsupervised classification is epitomized, and how bottom-up and autonomous concept acquisitions are accomplished.

This system acquires concepts autonomously from vast amounts of input signals that are influenced by noise. The ANG used for the bottom layer of this system can cope with on-line unsupervised learning and elimination of noise. Based on the result of learning by ANG, a layered structure is constructed autonomously through bottom-up conceptual acquisition. The purpose of this process is the preparation to cope with the difficulty in the ANG, the high-density overlaps.

In general, there are overlaps between clusters. In the ANG, it is presumed that the input data are separable, meaning that the probability density in the centric part of every cluster is higher than the density in the intermediate parts between clusters: the overlaps between clusters will have low probability density [7]. Nevertheless, input data which are acquired constantly in the real world are not separable. In the case of the feature vector of a color or a form, for example, it is certain that two or more concepts exist with almost the same color or form. On the other hand, it is also certain that concepts with various colors or forms exist. Therefore, when symbols are related with the result classified according to real-world data, some clusters may not correspond one-to-one with the symbols. In this system, a cluster which is constituted of ANG can be divided or be shared by multiple concepts.

In data analysis layers, the result of classification by the ANG is epitomized. About each cluster in the ANG layer, nodes with much number of times of approximation of an input, i.e., the nodes judged to be located in high probability density serve as a prototype. They are reproduced in the upper layer. Other nodes are connected with the nearest prototype. Consequently, each cluster is expressed by a smaller number of prototype in the upper layer. The prototype is attracted by the prototypes of the same class. It is repelled by the prototypes of a different class. It receives the force of stay from nodes in a lower layer, and changes its location by these effects. Then, the prototypes with almost the same position are combined to further simplify representation.

The number of the layers of a data analysis layer changes with the threshold of the probability density for reproduction of the node to the upper layer. Until each cluster is represented

by several prototypes, the data analysis layer continues to create an upper layer with a stronger threshold and more powerful forces.

In the concept acquisition layer, concepts are acquired based on the result of two or more data analysis layers. Information regarding the nearest node of the ANG layer which reacted to each input data is conveyed to the prototype of the upper layer which is connected to it. Then, based on the information of the prototype which reacted in each data analysis layer, the concept acquisition layer judges the input to be the concept acquired previously, or to be a new concept. That is, the process is both learning and recognition.

In this study, we presume that some signals of the same concept are input directly because, for the constant learning task in the real world, we can be fairly certain that two or more signals are acquired while having caught a certain target for learning. From this assumption, the information that a set of input is the input data obtained from the same target for learning is useful. The concept acquisition layer operates learning and recognition based on the prototype which reacted, and its number of reactions. When a new concept is acquired, it is connected to each prototype that reacted in each data analysis layer. Thereby, it utilizes the following input signals for recognition.

C. Algorithm for bottom-up concept acquisition

With the analysis of Section 2.2, we give an algorithm for bottom-up concept acquisition. In the following description, we define the structural analysis layer as comprising two layers.

1) Notations used in the algorithm:

- A_i is the prototype set of the data analysis layer i, which is used to store the prototypes.
- C is the concept set of the concept acquisition layer, which is used to store the concept nodes.
- L_i is the connection set of data analysis layer i, which is used to store the connections between concept node and prototypes.
- W_i is the *n*-dimensional weight vector of the node or the prototype i.
- M_i is the locally accumulated number of signals of node i; the number is updated when node *i* is the winner.
- CL_i is the cluster number. This will be used to judge which cluster node *i* in data analysis layer belongs to.
- G is the node set to store the nodes. It is reproduced to the upper layer.
- un_i is the prototype connected with node i.
- DN_i is the node set to store the nodes connected to prototype i.
- $age_{(i,j)}$ is the age of the connection that connects concept node i and prototype j.
- N_i is the prototype set to store the prototypes connected to the concept node i.

2) the ANG layer: The ANG layer operates local unsupervised classification and topology representation. It can be called local because it is the classification only for each space. When the Euclidean distance between the input and the nearest node is larger than a threshold distance T, the node is set as a first node of a new cluster. If the number of input signals generated up to that point is an integer multiple of λ_1, it determines whether a new node is inserted. Fundamental processing remains unchanged; therefore, we omit a detailed description here.

3) Data analysis layer except the ANG:

step 0. Initialize the prototype set A as empty set: $A = \emptyset$.

step 1. Information of the nearest node s in the ANG layer is transmitted. That is, if the related prototype exists, then it begins:$\xi = s$.

step 2. If the number of input signals generated to this point in the process is an integer multiple of a parameter λ_2, and if a node with the accumulated number of signals M which fills the following criteria exists, except where the node is connected a prototype already, create the new prototype u to A using the following procedures. If $M_c > M_{thre} \quad , \quad M_c/M_{sum} > M_{ratio} \quad (\forall c\{c|c \in ANG, un_c = \emptyset\}) \quad (M_{sum} = \sum_{CL_k=CL_c} M_k)$, then:

- An original node of the new prototype g is set as c. The weight vector and class number of the prototype are set identically to that of the original node. Insert the prototype to the prototype set A.

$$g = c \tag{1}$$
$$G = G \cup \{g\} \tag{2}$$
$$W_u = W_g \tag{3}$$
$$CL_u = CL_g \tag{4}$$
$$A = A \cup \{u\} \tag{5}$$

- Eliminate a previous connection of an original node and create the connection between a new node and an original node.

$$DN_{un_c} = DN_{un_c} \setminus \{c\} \tag{6}$$
$$un_c = u \tag{7}$$
$$DN_u = \{c\} \tag{8}$$

- If s is removed at a process of the ANG, the following steps are not performed: go to step 1.

step 3. If ξ is not connected to any prototypes, create a connection for ξ. Create the connection to the prototype corresponding to the nearest original node against ξ. If $un_\xi = \emptyset$, $A \neq \emptyset$, then $o = \arg\min_{c \in X} \parallel \xi - W_{g_c} \parallel (X = \{i|i \in A, CL_i = CL_\xi\})$, $DN_o = DN_o \cup \{\xi\}$, $un_\xi = o$.

step 4. If some edges, nodes, or prototypes are inserted or removed by the above-mentioned process, judge the necessity of change for the connections that are unrelated to the original nodes. If $c \notin G \ \ (\forall c \in ANG)$,

$g \in X \quad (\forall g \in G) \quad (X = \{i | i \in G, CL_i = CL_c\})$, then

$$p = \arg\min_{g \in X} \| W_c - W_g \| \tag{9}$$

$$DN_{un_c} = DN_{un_c} \setminus \{c\} \tag{10}$$

$$un_c = un_p \tag{11}$$

$$DN_{un_p} = DN_{un_p} \cup \{c\} \tag{12}$$

step 5. If at least one prototype exists, operate the following processes to calculate the forces concerning the prototype.

- Calculate the forces F_c concerning the prototype. F_c is expressed with the sum of $Ue(c)$, $Ud(c)$, and $D(c)$, where: $Ue(c)$ is an attractive force received from each prototypes of the same class; $Ud(c)$ is a repulsive force received from each prototype of different classes; and $D(c)$ is a force of stay from each nodes of the ANG layer. They are computed as follows: $o = UN_{s_1}$, $W_{oc} = W_o - W_c$. R_1 and R_2 are normal distribution functions. In addition, R_3 is a function with character that is opposite to R_1 and R_2.

$$F_c = Ue(c) + Ud(c) + D(c) \tag{13}$$

$$Ue(c) = \begin{cases} R_1(\| W_{oc} \|)W_{oc} \\ (\forall CL_c = CL_o) \\ \sum_{CL_d = CL_c} R_1(\| W_{cd} \|)W_{dc} \\ (\forall CL_c \neq CL_o) \end{cases} \tag{14}$$

$$Ud(c) = \sum_{CL_d \neq CL_c} R_2(\| W_{cd} \|)W_{cd} \tag{15}$$

$$D(c) = \sum_{d \in DN_c} R_3(\| W_{cd} \|)W_{cd} \tag{16}$$

- Update the weight vector of each prototype by the forces associated with them.

$$W_c = W_c + F_c \quad (\forall c \in A) \tag{17}$$

step 6. If s is connected with a certain prototype, increment the number of reactions of its prototype: $CR_c = CR_c + 1 \quad (c \in A, c = un_s)$.

step 7. Go to step 1 to continue epitomizing the result of the ANG layer.

4) Concept acquisition layer:

step 0. Initialize the concept set C, the accumulated number of input signals E_{input}, and the accumulated number of reactions: $C = \emptyset, E_{input} = 0, E_{reaction} = 0$.

step 1. Acquire the reacted prototype ξ_n from n data analysis layers. Then increment the accumulated number of input signals: $\xi_n = p_n$, $E_{input} = E_{input} + 1$.

step 2. If one or more concepts with the connections to all the reacted prototypes exist, increment the accumulated number of reactions and the number of reactions of all the concepts: $CR_c = CR_c + 1$ $(\forall c \in C)$, $E_{reaction} = E_{reaction} + 1$.

step 3. If the input data set is disrupted, if the accumulated number of input signals is an integer multiple of a parameter λ_3, or if the accumulated number of reactions is larger than a threshold E, choose the concept node q with most numerous reactions.

$$q = \arg\max_{c \in C} CR_c \tag{18}$$

step 4. If there is no concept node or if the number of reactions of the selected concept node is smaller than a threshold T_{CR}, create the new concept node r: if $C = \emptyset$ or $CR_q < T_{CR}$, then $N_r = \emptyset, C = C \cup \{r\}, q = r$.

step 5. If connections between q and the reacted prototypes of data analysis layer do not exist already, create them and add them to the connection set L. If they exist, set the age of the connections to zero: if $CR_c > T_{CR}$, then $L_i = L_i \cup \{(q, c)\}, Nq = Nq \cup \{c\}$ $(\forall c \in A_i)$; $age_{(q,c)} = 0 (\forall c \in A_i)$ $(1 \leq i \leq n)$.

step 6. According to the number of reactions of the reacted prototypes and q, increase the age of the connections which connect them. In addition, incrementally change the ages of all connections of q: $age_{(q,c)} = age_{(q,c)} + 1$, if $CR_c > T_{CR}$, then $age_{(q,c)} = age_{(q,c)} + CR_q - CRc (\forall c \in N_q)$.

step 7. Remove connections with an age greater than some threshold $link_{dead}$: if $(x, y) \in L_i$, $age_{(x,y)} > link_{dead} (\forall x, y \{x \in C, y \in A_i\})$, then $L_i = L_i \setminus \{(x, y)\} (1 \leq i \leq n)$.

step 8. If a concept node has one or fewer connections of each data analysis layer, remove the concept node and its connections: if $P_{x_i} < 1 (\forall x \in C)$, then $L_i = L_i \setminus \{(x, y)\} (\forall y \in N_x), C = C \setminus \{x\} (1 \leq i \leq n)$.

step 9. Reset all numbers of reactions of the concept nodes and the prototypes, the accumulated number of input signals, and the accumulated number of reactions. Then go to step 1 to continue acquisition of concept: $CR_c = 0 \quad (\forall c \in \{A \cup C\})$, $E_{input} = 0$, and $E_{reaction} = 0$.

III. Simulation

A. Setting about input data set

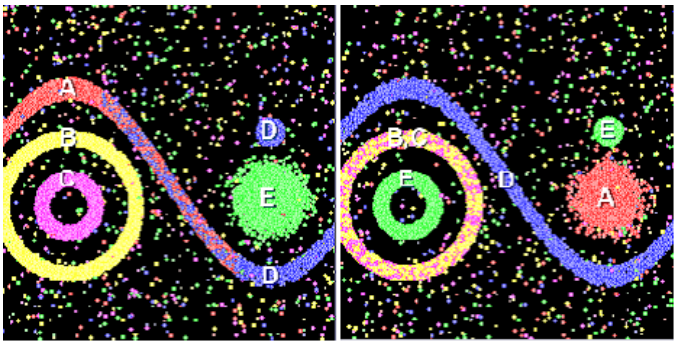

Fig. 2. Data set used for simulations: The five combinations of the input data distribution in two space are shown by the five letters.

We perform our simulation using the data set shown in Fig. 2. Two artificial two-dimensional data sets are used to take advantage of their intuitive manipulation, visualization and resulting insight into this system's function. The data set of Fig. 2 is separated to five parts: a sinusoidal shape, a famous single-link example, a circular area, and a data set satisfying a two-dimensional Gaussian distribution. The five combinations of the input data distribution in two space shown in Fig. 2 present some difficulties: B and C share a cluster; D and E have multiple clusters that occur in parallel; and A and D have a high-density overlap. We also add some random noise (5% of the useful data) to the data set to simulate real-world data. In such learning tasks, no previous techniques have been able to acquire suitable concepts.

In this simulation, we set the parameter $\lambda_1 = 100$, $\lambda_2 = 100$, $\lambda_3 = 100$, $age_{dead} = 50$, $link_{dead} = 400$, $M_{thre} = 500$, $M_{ratio} = 1.0$, $T_{merge} = 1.0$, $T_{CR} = 0.6$, $T_R = 5$, and $E = 0.3$.

B. Simulation in a stationary environment

Distribution of the input data is chosen randomly from the combination of A, B, C, D, and E. A set of patterns is chosen at random from the selected distributions. The input data set has an average of 100 signals. The progress to 250 000 signals of this simulation is presented in Fig. 3. The color of the node of the ANG layer represents the class into which the ANG is classified. Red squares represent prototypes of the data analysis layer. The prototype is connected with the related nodes of the ANG layer by yellow lines. The squares that line the base of a figure are the acquired concepts. The position of a concept in the figure is for intelligible display: it is not significant. The concept is connected with the related prototypes by the blue lines.

We next discuss results in detail. At the outset, the ANG of the bottom layer only performed unsupervised classification and topology learning. Subsequently, after about 50 000 signals, a prototype began to be generated. After about 100 000 signals, a combination of prototypes in two space came to exist. The concept was beginning its acquisition. Then, the concept was acquired gradually along with the increase in a prototype. After 250 000 signals, the proposed method was judged as having four classes: B and C, which share a cluster, were acquired as one concept; A and D, which have a high-density overlap, were acquired as discrete concepts that share some prototypes. This is the result of autonomous concept acquisition based on input unlabeled data. For that reason, it is not important in this simulation whether the distributions with overlap are judged in the same class or not. The result is appropriate in both cases and integration and separation of these concepts are problematic after getting input labeled data. Regarding both D and E, the multiple clusters that occur in parallel were acquired as one concept. Consequently, the system can acquire important concepts autonomously in the stationary environment. It will utilize limited labeled signals efficiently, and engender symbol grounding.

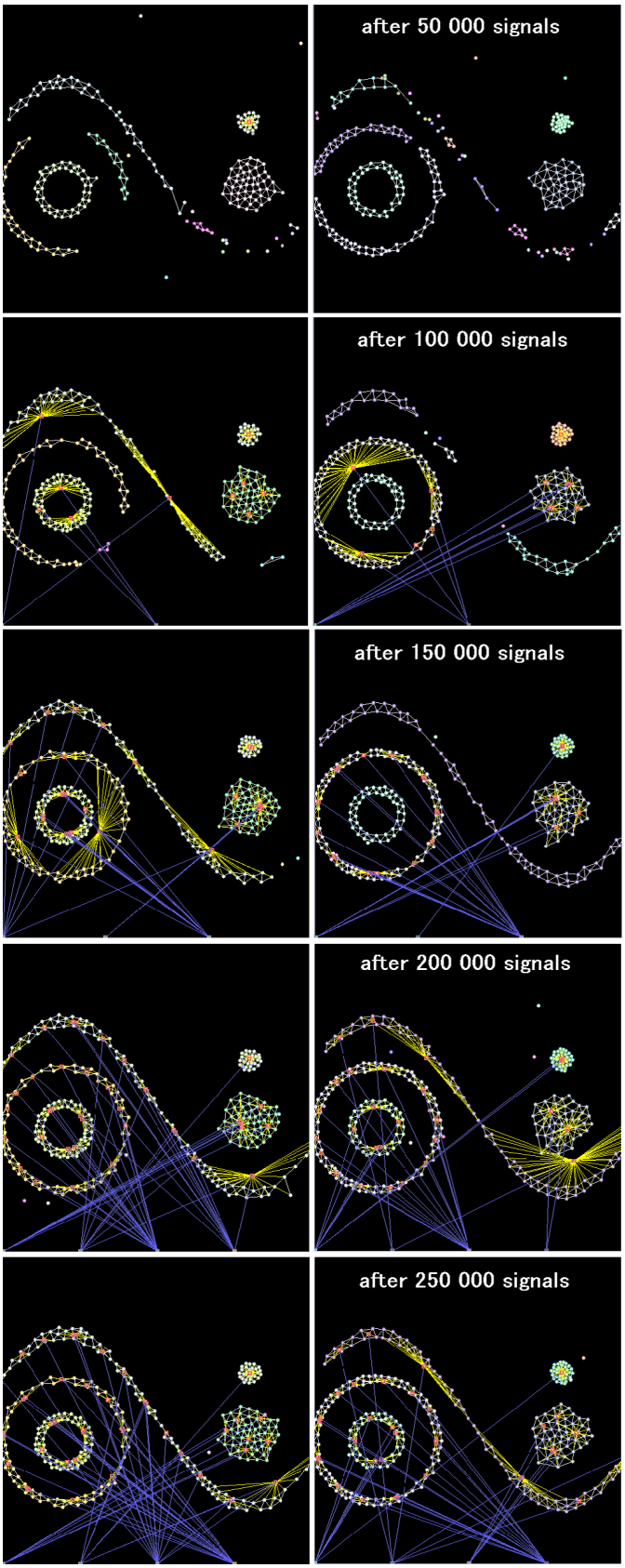

Fig. 3. Results of the concept acquisition in the stationary environment

C. Simulation in NON-stationary environment

Finally, we simulate the on-line concept acquisition process. The first 50 000 signals are chosen randomly from A. At signal 50 001, the environment changes and the 50 000 following signals are chosen from B. In a similar manner, respective groups of 50 000 signals are chosen from other sets in the order of C, D, and E. For every environment, we add 5% noise to the test data; the noises are distributed throughout the whole data space.

Figure 4 shows results of our proposed method. After learning under one environment, we report the intermediate conceptual structures. As the figure indicates, the concept was acquired sequentially after data input from A or B. The concept related to B and the prototypes of C were connected after data input from C. However, existing connections with the prototypes of B were deleted. After the data input from D, the new concept was acquired sharing the existing prototype of A. The new concept was acquired sequentially and satisfactory also about the input from E. Four concepts were acquired sequentially by the end of this simulation. This result was the same result as the simulation in the previously described stationary environment. However, on the analogy about the distribution in one of the two space, such as B and C, the result in this simulation is not suitable. This problem must be solved in the future to use limited input labeled data efficiently.

IV. Conclusion

This paper presented a new on-line learning and recognition method for hierarchization of data distributions and concept acquisition. The algorithm is able to constitute the layered structure of a data distribution based on inputs from two or more space; it can also acquire concepts autonomously. In that regard, it is able to cope with some difficulties on this occasion, such as high-density overlaps, never-seen inputs, vast amounts of input signals, and noise. The simulations demonstrate that the problem of sequential concept acquisition is solved. Future studies will examine the effective utilization of limited input labeled signals.

References

[1] B.Fritzke: *A growing neural gas network learns topologies*, Advances in Neural Information Processing Systems 7, MIT Press, pp.625–632, 1995.

[2] J.Bruske and G.Sommer: *Dynamic cell structure learns perfectly topology preserving map*, Neural Computation, Vol.7, pp.845–865, 1995.

[3] F.Shen and O.Hasegawa: *A growing neural network for online unsupervised learning*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.8, No.2, pp.121–129, 2004.

[4] G.A.Carpenter and S.Grossberg: *The ART of adaptive pattern recognition by a self organizing neural network*, IEEE Computer, Vol.21, pp.77–88, 1988.

[5] B.Fritzke: *A self-organizing network that can follow non-stationary distributions*, In Proceedings of ICANN-97, pp.613–618, 1997.

[6] C.P.Lim and R.F.Harrision: *An incremental adaptive network for on-line supervised learning and probability estimation*, Neural Networks, Vol.10, pp.925–939, 1997.

[7] F.Shen and O.Hasegawa: *A growing neural network for online unsupervised learning*, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.8, No.2, pp.121–129, 2004.

[8] Y.Kamiya, et al.: *A proposal of non-stop learning and recognition scheme for real-world pattern information*, The 18th Annual Conference of JSAI, 2004. (In Japanese).

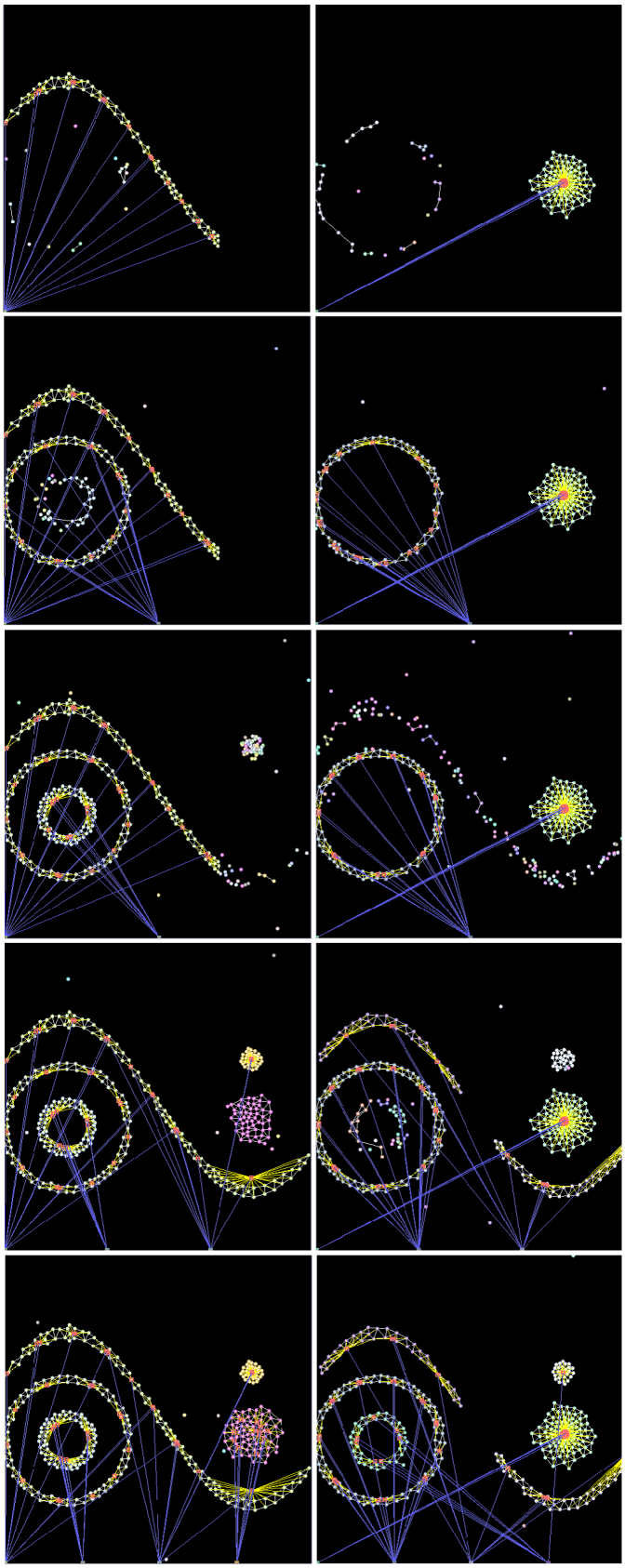

Fig. 4. Results of sequential concept acquisition in a non-stationary environment